

ACCURATE AND RELIABLE CANCER CLASSIFICATION BASED ON
PATHWAY-MARKERS AND SUBNETWORK-MARKERS

A Thesis

by

JUNJIE SU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

December 2010

Major Subject: Electrical Engineering

ACCURATE AND RELIABLE CANCER CLASSIFICATION BASED ON
PATHWAY-MARKERS AND SUBNETWORK-MARKERS

A Thesis

by

JUNJIE SU

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE

Approved by:

Chair of Committee,	Byung-Jun Yoon
Committee Members,	Edward R. Dougherty
	Ulisses Braga Neto
	Robert Chapkin
	Shuguang (Robert) Cui
Head of Department,	Costas N. Georgiades

December 2010

Major Subject: Electrical Engineering

ABSTRACT

Accurate and Reliable Cancer Classification Based on Pathway-Markers and
Subnetwork-Markers. (December 2010)

Junjie Su, B.E., Tsinghua University

Chair of Advisory Committee: Dr. Byung-Jun Yoon

Finding reliable gene markers for accurate disease classification is very challenging due to a number of reasons, including the small sample size of typical clinical data, high noise in gene expression measurements, and the heterogeneity across patients. In fact, gene markers identified in independent studies often do not coincide with each other, suggesting that many of the predicted markers may have no biological significance and may be simply artifacts of the analyzed dataset. To find more reliable and reproducible diagnostic markers, several studies proposed to analyze the gene expression data at the level of groups of functionally related genes, such as pathways. Given a set of known pathways, these methods estimate the activity level of each pathway by summarizing the expression values of its member genes and using the pathway activities for classification. One practical problem of the pathway-based approach is the limited coverage of genes by currently known pathways. As a result, potentially important genes that play critical roles in cancer development may be excluded. In this thesis, we first propose a probabilistic model to infer pathway/subnetwork activities. After that, we developed a novel method for identifying reliable subnetwork markers in a human protein-protein interaction (PPI) network based on probabilistic inference of subnetwork activities. We tested the proposed methods based on two independent breast cancer datasets. The proposed method can efficiently find reliable subnetwork markers that outperform the gene-based and pathway-based markers in

terms of discriminative power, reproducibility and classification performance. The identified subnetwork markers are highly enriched in common GO terms, and they can more accurately classify breast cancer metastasis compared to markers found by a previous method.

ACKNOWLEDGMENTS

First, I would like to thank all my committee members, especially my advisor, Byung-Jun Yoon, and Edward Dougherty, for their help on my research. I also would like to thank the authors of "Network-based Classification of Breast Cancer Metastasis", especially H.Y. Chuang and T. Ideker, for sharing the PPI network and their helpful communication.

TABLE OF CONTENTS

CHAPTER		Page
I	INTRODUCTION	1
II	ACCURATE AND RELIABLE CANCER CLASSIFICATION BASED ON PROBABILISTIC INFERENCE OF PATHWAY ACTIVITY	3
	A. Methods	4
	1. Datasets	4
	2. Probabilistic Inference of Pathway Activity	4
	3. Discriminative Power of Pathway Markers	6
	4. Evaluation of Classification Performance	8
	5. Computing the Area Under ROC Curve	11
	B. Results	12
	1. Probabilistic Pathway Activity Inference Improves the Discriminative Power of Pathway Markers	12
	2. Proposed Pathway Activity Inference Scheme Leads to More Accurate and Reliable Classifiers	15
	3. Proposed Method Leads to Robust Classifiers that Yield Symmetric Results for Dataset Inversion	21
	C. Discussion	22
III	IDENTIFICATION OF DIAGNOSTIC SUBNETWORK MARK- ERS FOR CANCER IN HUMAN PROTEIN-PROTEIN IN- TERACTION NETWORK	26
	A. Results and Discussion	26
	1. Identification of Subnetwork Markers	26
	2. The Identified Subnetworks are Enriched with Pro- teins in Common GO Terms	30
	3. Subnetwork Markers Identified by the Proposed Method are More Discriminative and Reproducible	30
	4. Subnetwork Markers Identified by the Proposed Method Improves Classification Performance	36
	B. Conclusions	42
	C. Methods	43

CHAPTER		Page
	1. Overview	43
	2. Probabilistic Inference of Subnetwork Activity	45
	3. Evaluating the Discriminative Power of Linear Paths in the PPI Network	45
	4. Searching for Discriminative Linear Paths	46
	5. Combining Top Overlapping Paths into a Subnetwork	48
IV	GENE CLUSTERING IN THE PPI NETWORK BASED ON A MESSAGE PASSING ALGORITHM TOWARDS AC- CURATE DISEASE CLASSIFICATION	49
	A. Methods	50
	1. Affinity Propagation	50
	2. Computing the Similarities between Genes	51
	B. Results and Discussion	53
	1. The Identified Gene Clusters Improve Classifica- tion Performance Significantly	53
	2. Affinity Propagation Finds Better Combination of Genes	57
	C. Conclusions	57
V	CONCLUSION	60
	REFERENCES	61
	VITA	67

LIST OF TABLES

TABLE		Page
I	Statistics of the subnetwork markers identified by the proposed method.	29
II	Enrichment analysis results for the sample subnetworks shown in Figure 8.	32
III	Average size of the identified gene clusters.	54
IV	Statistics of the gene clusters identified using affinity propagation and the subnetworks identified using the method proposed in Chapter III.	58

LIST OF FIGURES

FIGURE		Page
1	Probabilistic inference of pathway activity.	7
2	Illustration of the experimental set-up.	11
3	Discriminative power of prescreened pathway markers and single gene markers.	14
4	Discriminative power of all pathway markers and gene markers.	16
5	Performance of different classification methods.	19
6	Performance of different classification methods.	20
7	Robustness of the proposed classification scheme.	22
8	Sample subnetworks identified using the proposed method.	31
9	Discriminative power of the subnetwork markers identified by the proposed method using different θ	35
10	Discriminative power of different types of markers.	37
11	Classification performance of the identified subnetwork markers for different θ	39
12	Classification performance of different types of markers.	40
13	Classification error at different TPR (true positive rate) for different types of markers.	41
14	Illustration of the proposed method.	44
15	Classification performance of gene clusters identified using different α . . .	55
16	Statistics of the gene clusters identified using different α	56
17	Discriminative power of the identified gene clusters and subnetworks. . . .	58

CHAPTER I

INTRODUCTION

The introduction of affordable microarray technologies for measuring genome-wide expression profiles has led to the development of numerous methods for discriminating between different classes of a complex disease, such as cancer, through transcriptome analysis [1, 2, 3, 4]. Especially, there have been significant research efforts to identify differentially expressed genes across different phenotypes [5, 6, 7, 8, 9], which can be used as diagnostic markers for classifying the disease states or predicting the outcome of medical treatments [1, 2, 3, 4, 10, 11, 12]. However, finding reliable gene markers is a challenging problem, and several recent studies have questioned the reliability of many classifiers based on individual gene markers [13, 14, 15, 16, 17, 18, 19]. The small sample size of typical clinical data that are used to build a classifier is one of the major factors that make this problem difficult. We often have to search for a small number of good marker genes among thousands of genes based on a limited number of samples, which makes the performance of traditional feature selection methods quite unpredictable [20]. The inherent measurement noise in high-throughput experimental data and the heterogeneity across samples and patients make the problem even more formidable.

One possible way to address this problem is to interpret the expression data at the level of functional modules, such as signaling pathways and molecular complexes, instead of at the level of individual genes. In fact, one of the weaknesses of many gene-based classification methods is that the marker genes are often selected independently, even though their functional products may interact with each other.

The journal model is *IEEE Transactions on Automatic Control*.

Therefore, the selected gene markers may contain redundant information, and they may not synergistically improve the overall classification performance. We can alleviate this problem by jointly analyzing the expression levels of groups of functionally related genes, which can be obtained based on transcriptome analysis [21, 22, 23], GO annotations [24], or other sources. In fact, several studies [23, 25, 26, 27, 28] have shown that pathway markers are more reproducible compared to single gene markers and they can provide important biological insights into the underlying mechanisms that lead to different disease phenotypes. Furthermore, pathway-based classifiers often achieve comparable or better classification performance compared to traditional gene-based classifiers.

Pathway-based methods also have some shortcomings. First, currently known pathways cover only a limited number of genes and may not include key genes with significant expression changes across different phenotypes. Besides, many pathways overlap with each other, hence the activity of such pathway markers may be highly correlated.

To alleviate these problems, one way is to directly identify such markers in a large protein-protein interaction (PPI) network. In a recently published paper [29], Chuang et al. tried to identify subnetwork markers by overlaying gene expression data on the corresponding proteins in a PPI network. They started from the so-called seed proteins in the PPI network that have high discriminative power and greedily grew subnetworks from them to maximize the mutual information between the subnetwork activity score and the class label. They showed that subnetwork markers yield more accurate classification results and have better reproducibility compared to gene markers.

CHAPTER II

ACCURATE AND RELIABLE CANCER CLASSIFICATION BASED ON
PROBABILISTIC INFERENCE OF PATHWAY ACTIVITY*

To use pathway-based markers in classification, we need a way to infer the activity of a given pathway based on the expression levels of the constituent genes. Recently, a number of pathway activity inference methods have been proposed for this purpose. For example, Guo et al. [25] proposed to use the mean or median expression value of the member genes to infer the pathway activity. Tomfohr et al. [28] and Bild et al. [23] used the first principal component of the expression profile of the member genes to estimate the activity of a given pathway. More recently, Lee et al. [26] proposed a method that predicts the pathway activity using only a subset of genes in the pathway, called the condition-responsive genes (CORGs), whose combined expression levels can accurately discriminate the phenotypes of interest.

In this chapter, we propose a novel method for probabilistic inference of pathway activities. For a given pathway, the proposed method estimates the log-likelihood ratio between different phenotypes based on the expression level of each member gene. The activity level of the pathway is then inferred by combining the log-likelihood ratios of the genes that belong to the pathway. We apply our method to the classification of breast cancer metastasis, and demonstrate that it can achieve higher accuracy compared to several previous pathway-based approaches. Furthermore, we show that the proposed pathway activity inference method can find more reproducible pathway markers that retain the discriminative power across different datasets.

*Reprinted with permission from “Accurate and Reliable Cancer Classification based on Probabilistic Inference of Pathway Activity” by J. Su, B.J. Yoon and E.R. Dougherty, *PLOS ONE*, vol. 4, pp. e8161, 2009. Copyright 2010 by PLOS ONE.

A. Methods

1. Datasets

We obtained two independent breast cancer datasets from large-scale gene expression studies by Wang et al. [11] (referred as the USA dataset in this work) and van't Veer et al. [10] (referred as the Netherlands dataset). Wang et al.'s dataset [11] contains the gene expression profiles of 286 breast cancer patients from the USA, where metastasis was detected in 107 of them while the remaining 179 were metastasis-free. The other dataset studied by van't Veer et al. [10] contains the gene expression profiles of 295 patients from the Netherlands, where 79 had metastasis and 216 were metastasis-free. In this study, we did not consider the follow-up time or the occurrence of distant metastasis.

To obtain the set of known biological pathways, we referred to the MSigDB (Molecular Signatures Database) version 2.4 (updated April 7, 2008) [21]. We downloaded the canonical pathways in the C2 curated gene sets, which contains 639 gene sets obtained from several pathway databases, including the KEGG (Kyoto Encyclopedia of Genes and Genomes) database [30] and the GenMAPP [31]. These gene sets are compiled by domain experts and they provide canonical representations of biological processes. The set of pathways obtained from the MSigDB covers more than 5,000 distinct genes, where 3,271 of them can be found in both microarray platforms used by the two breast cancer gene expression studies in [10, 11].

2. Probabilistic Inference of Pathway Activity

For each pathway, we first identified the genes that were included in the expression profiles in the two breast cancer datasets. The genes that were not included in these datasets were removed from the gene set for the given pathway. Consider a pathway

that contains n genes $\mathcal{G} = \{g_1, g_2, \dots, g_n\}$ after removing the genes whose expression values were not available. Given a sample $x_j = (x_j^1, x_j^2, \dots, x_j^n)$ that contains the expression levels of the member genes, we estimate the pathway activity a_j as follows

$$a_j = \sum_{i=1}^n \lambda_i(x_j^i), \quad (2.1)$$

where $\lambda_i(x_j^i)$ is the log-likelihood ratio (LLR) between the two phenotypes of interest for the gene g_i . The LLR $\lambda_i(x_j^i)$ is given by

$$\lambda_i(x_j^i) = \log[f_i^1(x_j^i)/f_i^2(x_j^i)], \quad (2.2)$$

where $f_i^1(x)$ is the conditional probability density function (PDF) of the expression level of gene g_i under phenotype 1, and $f_i^2(x)$ is the conditional PDF under phenotype 2. The ratio $\lambda_i(x_j^i)$ is a probabilistic indicator that tells us which phenotype is more likely based on the expression level x_j^i of the i th member gene g_i . We combine the evidence from all the member genes to infer the overall pathway activity $a_j = \sum_{i=1}^n \lambda_i(x_j^i)$. The pathway activity a_j can serve as a discriminative score for classifying the sample x_j into different phenotypes based on the activation level of the given pathway. Conceptually, we can view this approach as computing the relative support for the two different phenotypes using a Naive Bayes model [31, 32] based on the gene expression profile of the pathway.

In order to compute the LLR value $\lambda_i(x_j^i)$, we need to estimate the PDF $f_i^c(x)$ for each phenotype $c \in \{1, 2\}$. We assume that the gene expression level of gene g_i under phenotype c follows a Gaussian distribution with mean μ_i^c and standard deviation σ_i^c . These parameters were estimated based on all available samples x_j^i that correspond to the phenotype c . The estimated PDFs can then be used for computing the log-likelihood ratios. In practical applications, we often do not have enough training data for reliable estimation of the PDFs $f_i^1(x)$ and $f_i^2(x)$. This may make

the computation of LLRs sensitive to small changes in the gene expression profile. To avoid this problem, we normalize the $\lambda_i(x_j^i)$ as follows

$$\hat{\lambda}_i(x_j^i) = \frac{\lambda_i(x_j^i) - \mu(\lambda_i)}{\sigma(\lambda_i)}, \quad (2.3)$$

where $\mu(\lambda_i)$ and $\sigma(\lambda_i)$ are the mean and standard deviation of $\lambda_i(x_j^i)$ across all samples, respectively. Figure 1 illustrates the overall procedure for inferring the activity of a given pathway.

3. Discriminative Power of Pathway Markers

In order to compare the proposed pathway activity inference scheme with other existing methods, we performed the following experiments. In our first experiment, we selected the top 50 differentially expressed pathways using the method proposed by Tian et al. [22]. To assess the ability of a given pathway in discriminating between different phenotypes, Tian et al. computes the *t*-test statistics scores for all member genes and take their average to compute an aggregated score T that can serve as an indicator of the pathways discriminative power. After prescreening the top 50 pathways that have the largest absolute T values, we computed the activity score for each of these pathways using the proposed inference method as well as other methods. The obtained pathway activity scores were then used to compute the *t*-test statistics score for each pathway marker. The *t*-test scores were used to assess the discriminative power of pathway markers and to compare different inference methods.

In this work, we compared five different pathway activity inference methods: the mean and the median methods [25], the PCA-based method [23, 28], the CORG-based method [26], and the inference method proposed in this paper. For the mean, median, and CORG-based methods, we computed the score T by averaging the *t*-test scores of the expression values of the member genes. For the PCA-based method,

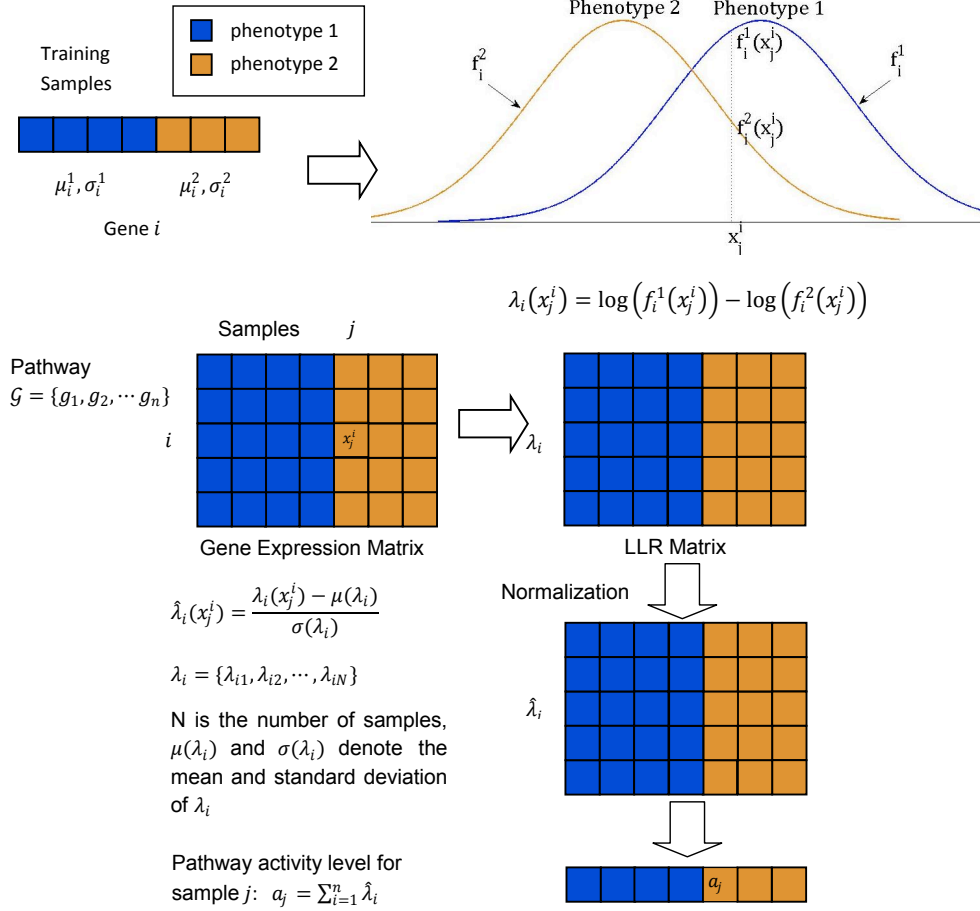


Fig. 1. Probabilistic inference of pathway activity.

For each gene in the pathway, we estimate the conditional probability density functions (PDFs) under different phenotypes. Based on the estimated PDFs, we transform the expression values of the member genes into log-likelihood ratios (LLRs) to obtain a LLR matrix from the gene expression matrix. The LLR matrix is then normalized, and the pathway activity is inferred by combining the normalized LLRs of its member genes.

we computed T by averaging the absolute t -test scores of the gene expression values, since the PCA can naturally combine expression values regardless of whether they are positively correlated or negatively correlated with the phenotype of interest. For our proposed method, we computed T by averaging the t -test scores of the LLRs of the member genes, since we estimated the pathway activity score based on LLRs instead of the original expression values.

We also evaluated the robustness of each inference method in identifying good pathway markers, by ranking the pathways using one of the two breast cancer datasets, and then assessing the discriminative power of the pathways based on the other dataset. Again, t -test statistics of the pathway activity scores were used to compare different inference methods.

In our second experiment, we computed the t -test statistics scores for all 639 pathways without any prescreening, and compared the effectiveness of different pathway activity inference methods based on the computed scores. As in the first experiment, we also evaluated the robustness of each inference method for finding effective pathway markers, by ranking the pathways according to the t -test scores estimated using one of the datasets, and then evaluating their discriminative power on the other dataset.

4. Evaluation of Classification Performance

In order to evaluate the classification performance of the proposed pathway activity inference method, we performed the following cross-validation experiments.

For within-dataset experiments, the samples in a dataset were randomly divided into five subsets of equal size, where the samples in four of these subsets were used for training the classifier and the remaining subset was used for assessing the classification performance. This has been repeated by using each subset as the test set to

obtain more reliable results. The training set was divided again into three equal-sized subsets. Two thirds were used for ranking the pathway markers and building the classifier (the marker-evaluation dataset), and one third of the training set was used for feature selection (the feature-selection dataset). All samples in the training set were used to estimate the PDFs of the gene expression values under different phenotypes. To build the classifier, we evaluated each pathway based on the discriminative power of its activity score to classify samples. The pathways were sorted in increasing order of the p-value. After ranking the pathways, we built the classifier, either based on logistic regression or LDA (linear discriminant analysis), as follows. Based on the marker-evaluation dataset, we first constructed the classifier with only one feature, namely, the pathway marker with the lowest p-value. The performance of the classifier was then measured by computing the AUC (Area Under ROC Curve) [32] on the feature-selection dataset. Next, we enlarged the set of features by selecting the pathway marker with the lowest p-value among the remaining pathways. A new classifier was trained using the selected features on the marker-evaluation dataset and its classification performance was again assessed on the feature-selection dataset. The added pathway marker was kept in the feature set if the AUC increased, and it was removed otherwise. We repeated the above process for all pathway markers to optimize the classifier. The performance of the optimized classifier was evaluated by computing the AUC on the test dataset. These experiments have been repeated for 100 random partitions of the entire dataset. We report the AUC, averaged over 500 experiments, as the overall performance measure of the classification method at hand. The overall process of the within-dataset experiment is illustrated in Fig. 2A.

In order to evaluate the reproducibility of the pathway markers across different dataset, we performed cross-dataset experiments, where one dataset was used for selecting the pathway markers, and the other dataset was used for building the classifier

based on the selected markers and evaluating its performance. First, we selected the optimal set of features (i.e., pathway markers) based on one dataset, by optimizing the AUC metric. The process for selecting the feature set was similar to the one used in the within-dataset experiments. The samples in the other dataset were divided into five subsets of equal size. Four fifths of samples were used to train the classifier using the selected features, and one fifth of samples were used to evaluate the performance of the constructed classifier. We repeated this experiment by using each of the five subsets as the test set and using the rest for training. The above experiment was repeated for 100 random partitions of the entire dataset, and the average AUC over the 500 experiments was reported as the performance measure. It is important to note that feature selection is performed solely based on the rst dataset. During the cross-validation experiments using the second dataset, the training set (that consists of four fifths of samples in the same dataset) is simply used to build the classifier based on the preselected set of features. The overall goal of these cross-dataset experiments is to evaluate the reproducibility of the feature set, selected using the proposed pathway activity inference scheme, across different datasets. Figure 2B illustrates the overall process of the cross-dataset experiment.

To compare the proposed method with other existing methods, we performed the described within-dataset experiments and the cross-dataset experiments using other pathway activity inference methods (mean, median, PCA, and CORG). In addition, we also evaluated the performance of a gene-based classifier that uses individual genes as diagnostic markers, following a similar procedure. In this study, we included the top 50 pathway markers in the initial marker set, which were selected according to the method in Tian et al. [22] as elaborated in the previous subsection. For the gene-based classifier, we included the top 50 gene markers with the lowest p-values in the initial marker set, in order to keep the maximum number of features identical.

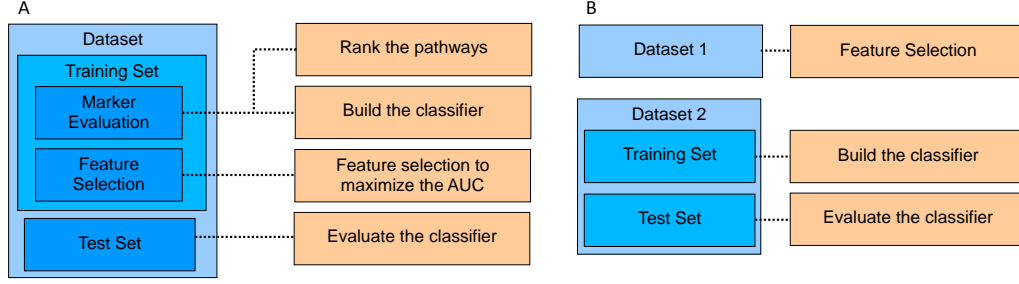


Fig. 2. Illustration of the experimental set-up.

(A) In the within-dataset experiments, part of the training set, referred as the marker-evaluation set, is used for ranking the pathway markers according to their discriminative power and building the classifier. The optimal set of features are selected based on the remainder of the training set, referred as the feature-selection set. The performance of the resulting classifier is evaluated using the test dataset. (B) In the cross-dataset experiments, one of the datasets is used to find the optimal set of features, and the other dataset is used to build a classifier based on the preselected features and to evaluate the classifier.

5. Computing the Area Under ROC Curve

In this work, we evaluated the performance of a classifier based on the AUC (Area Under ROC Curve). The AUC metric has been widely used for evaluating classification methods, since it can provide a useful summary statistics of the classification performance over the entire range of specificity and sensitivity values. To compute the AUC, we adopted the method proposed in [32]. For a given classifier, let x_1, x_2, \dots, x_m be the output of the classifier for positive samples, and let y_1, y_2, \dots, y_n be the output for negative samples. Then, the AUC metric A for the classifier is given by:

$$A = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n I(x_i > y_j), \quad (2.4)$$

where $I(\cdot)$ is the indicator function. The AUC is actually the empirical probability that a randomly chosen positive sample is ranked higher than a randomly chosen negative sample. It can be shown that the AUC measure is equivalent to the Mann-Whitney U -test (also called the Wilcoxon rank-sum test) statistics.

B. Results

1. Probabilistic Pathway Activity Inference Improves the Discriminative Power of Pathway Markers

We evaluated the discriminative power of pathway markers, where the pathway activities were inferred using the proposed method as well as other inference methods. For effective comparison of the proposed inference method with other existing methods, we carried out similar experiments as those performed in [26] to assess the discriminative power of pathway markers. For each breast cancer dataset, we first used the method of Tian et al. [22] to select the top 50 pathways among the 639 pathways obtained from the MSigDB [21] (see Methods). We computed the actual activity scores of the top 50 pathways based on each pathway activity inference scheme, and ranked the pathways according to their discriminative power. Figure 3 shows the discriminative power of the top pathways, where the x -axis corresponds to the number k of top pathways that were considered, and the y -axis shows the mean absolute t -score of the top k pathways. We compared five pathway activity inference methods, namely, the CORG- based method [26], PCA-based method [23, 28], mean and median methods [25], and the LLR-based method proposed in this paper. For comparison, we also evaluated the discriminative power of the top 50 single gene markers, which were chosen among the 3,271 genes covered by the 639 pathways used in this study. The results obtained from the Netherlands breast cancer dataset [10] and the USA breast cancer dataset [11] are shown in Fig. 3A and Fig. 3B, respectively. As we can see from these results, the proposed pathway activity inference scheme, which computes the pathway activity score by combining the log-likelihood ratios of the member genes, significantly improved the power of pathway markers to discriminate between metastatic samples and non-metastatic samples. Interestingly, the top gene

markers often compared favorably to pathway markers. On the Netherlands dataset, the expression levels of the top genes had larger discriminative power than the pathway activity scores inferred by the CORG, PCA, mean, and median methods. Only the pathway activity scores estimated by the proposed method were more discriminative than the gene expression values. On the USA dataset, gene markers were more discriminative than pathway markers based on mean, median, and PCA methods, but less discriminative compared to pathway markers based on the proposed method and the CORG method.

To evaluate the reproducibility of pathway markers, we ranked the markers based on one dataset and evaluated their mean absolute t -score using the other dataset. Figure 3C shows the result for ranking the markers based on the Netherlands dataset and computing the mean absolute t -score of the top k markers using the USA dataset. Similarly, Fig. 3D shows the result for ranking the markers based on the USA dataset and computing the mean score of the top k pathways using the Netherlands dataset. These results clearly show that the pathway markers selected based on the proposed inference method retain significantly large discriminative power across different datasets. In fact, in both cross-dataset experiments, the pathway activity scores computed by the LLR method were much more discriminative than the activity scores computed by other inference methods as well as the expression values of the top gene markers. Altogether, these results imply that the proposed method can find better diagnostic markers with higher reproducibility. Also note that the single gene markers, which had considerably large discriminative power within a dataset (see Figs. 3A and 3B), lost most of the discriminative power in a different dataset.

Next, we performed similar experiments for all 639 pathways and all 3,271 genes covered by these pathways, without any prescreening (see Methods). The results of these experiments are shown in Fig. 4, where the x -axis indicates the ratio $P\%$

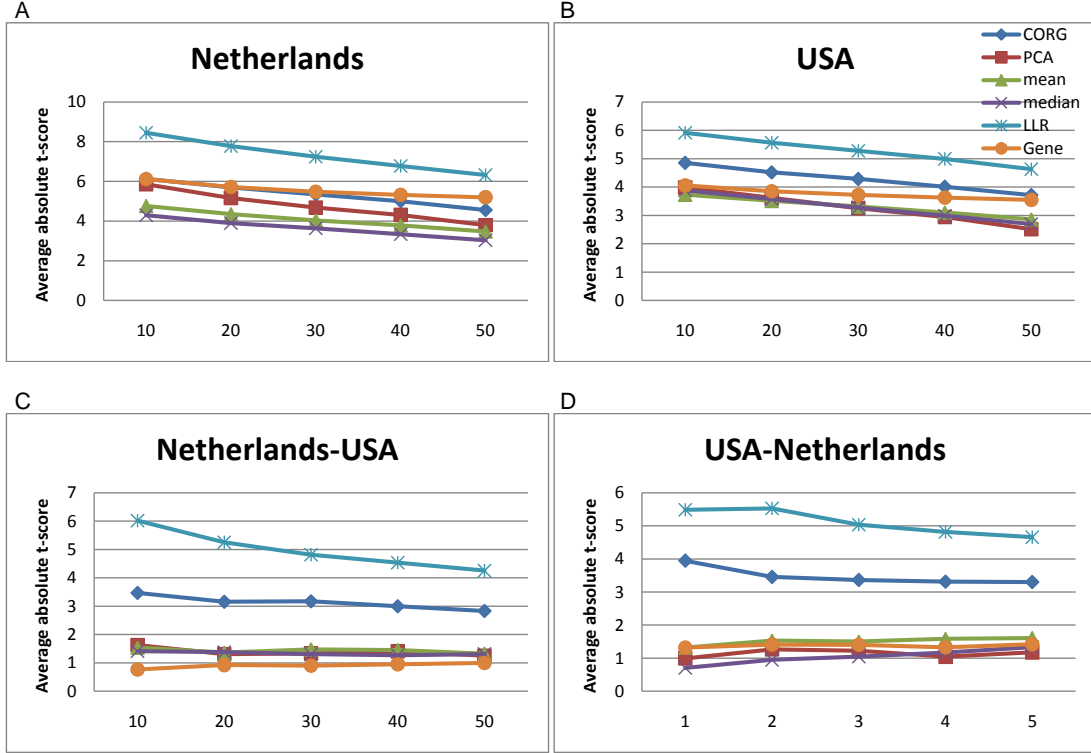


Fig. 3. Discriminative power of prescreened pathway markers and single gene markers.

(A) Mean absolute t -score of the top k ($k = 10, 20, 30, 40, 50$) markers for the Netherlands breast cancer dataset. Pathway activities have been inferred using five different methods: CORG, PCA, mean, median, and LLR (proposed method). The discriminative power of the top gene markers was estimated for comparison (labeled as Gene). (B) Mean absolute t -score of the top markers for the USA breast cancer dataset. (C) The markers were ranked based on the Netherlands dataset and the mean absolute t -score of the top k markers was computed based on the USA dataset. (D) The markers were ranked based on the USA dataset and the mean absolute t -score of the top markers was computed based on the Netherlands dataset.

of the top pathways that were used to compute the mean absolute t -score, and the y -axis corresponds to the estimated mean absolute t -score of the top $P\%$ pathways. The discriminative power of the pathway markers and the single gene markers on the Netherlands dataset is shown in Fig. 4A, and the discriminative power of the markers on the USA dataset is shown in Fig. 4B. The results obtained from cross-dataset experiments are summarized in Fig. 4C and 4D. In Fig. 4C, the markers were ranked according to their discriminative power on the Netherlands set, and their mean absolute t -scores were computed using the USA dataset. The results for ranking the markers based on the USA dataset and computing the scores using the Netherlands set are shown in Fig. 4D. All these experiments show that the pathway activity scores measured by the proposed LLR method are much more discriminative than the scores computed by other inference methods and also the expression values of individual genes. Furthermore, we can see that the pathway markers that were chosen based on the LLR-based pathway activity scores are more reproducible and their activity scores retain significant amount of discriminative capability across independent datasets.

2. Proposed Pathway Activity Inference Scheme Leads to More Accurate and Reliable Classifiers

We used the proposed pathway activity inference scheme for classification of breast cancer metastasis, to evaluate its usefulness in discriminating different cancer phenotypes. For a fair and effective comparison with other inference schemes, we again adopted a similar experimental set-up that was used in [26] to evaluate the performance of the CORG-based method, a state-of-the-art pathway activity inference scheme that uses only the condition-responsive genes in a given pathway. For each breast cancer dataset, we performed five-fold cross-validation experiments, where four fifths of samples were used for constructing the classifier and the remaining one fifth

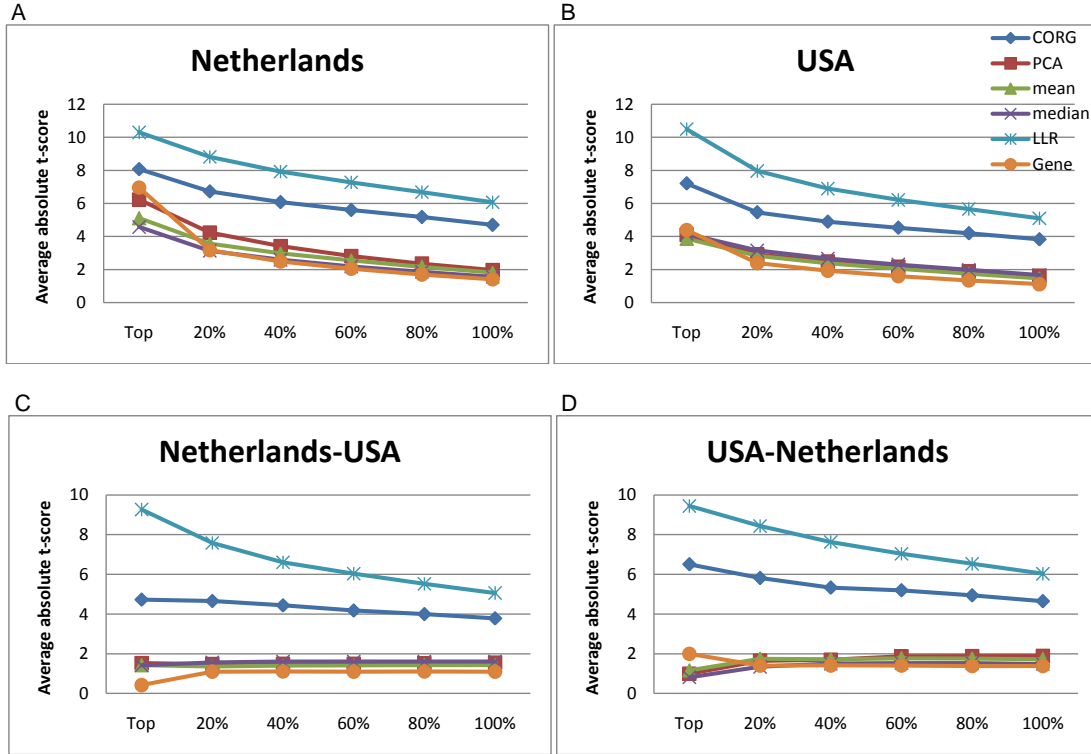


Fig. 4. Discriminative power of all pathway markers and gene markers.

(A) Mean absolute t -score of the top $P\%$ markers for the Netherlands dataset. (B) Mean absolute t -score of the top markers for the USA dataset. (C) The markers were ranked based on the Netherlands dataset and the mean absolute t -score of the top $P\%$ markers was computed based on the USA dataset. (D) The markers were ranked based on the USA dataset and the mean score of the top $P\%$ markers was computed based on the Netherlands dataset.

of samples were used for evaluating the classification performance (see Methods). While constructing the classifier, we used the LLR-based pathway activity inference method for assessing the discriminative power of each pathway marker and selecting the optimal set of markers to be used in the classifier. The constructed classifier also used the pathway activity scores computed by the proposed inference method to distinguish metastatic breast cancer samples from non- metastatic samples. In our experiments, we defined the initial set of pathway markers as the top 50 pathways selected using the method by Tian et al. [22] (see Methods). We assessed the classification performance using the AUC metric. We repeated the five-fold cross-validation for 100 random partitions of the given dataset, and averaged the resulting 500 AUCs to obtain a reliable performance measure of the classification method. To compare the classification performance of different inference methods, we also repeated the previous experiments using the CORG, PCA, mean, and median methods for inferring the pathway activities. For comparison, we also evaluated the performance of the gene-based classification method. We included the top 50 discriminative genes in the initial marker set, to keep the maximum number of features identical for all classification methods.

Figure 5 summarizes the results of the cross-validation experiments. In the first set of experiments, we used logistic regression for classifying the samples. The classification results of different approaches based on logistic regression are shown in Fig. 5A. The two bar charts on the left of Fig. 5 correspond to the two within-dataset experiments based on the USA breast cancer dataset (labeled as USA) and the Netherlands dataset (labeled as Netherlands), respectively. In these within-dataset experiments, the initial set of top 50 markers have been selected using the entire dataset, in order to reduce the effect of sensitivity in marker selection when comparing different pathway-based methods. The cross-validation experiments have been performed based on the

selected initial set of markers (see Methods). As we can see in these bar charts, the proposed method achieved the highest classification accuracy among all methods, in both experiments. The CORG-based method compared favorably to other pathway-based methods, though outperformed by the proposed method. We can also see that the gene-based classifier performed very well in within dataset experiments, which is not surprising if we consider the high discriminative power of the top gene markers observed in Figs. 3A and 3B.

The results of the cross-dataset experiments are shown in the two bar charts on the right of Fig. 5A. The chart labeled as USA-Netherlands shows the results for selecting the features using the USA dataset, and training/evaluating the classifier using the Netherlands dataset. Similarly, the chart labeled as Netherlands-USA shows the classification performance for choosing the feature set using the Netherlands dataset, and training and evaluating the classifier based on the USA dataset. As we can see, the proposed LLR-based method outperformed most of the other methods in both cross-dataset experiments. Only the mean-based approach showed better performance than the proposed approach on the Netherlands-USA cross-dataset experiment. These results show that the proposed pathway activity inference method can find a better feature set that is more reproducible across datasets, compared to other activity inference methods. Despite the good performance in within-dataset experiments, gene-based classifiers performed typically worse than many pathway-based classifiers, which shows the poor reproducibility of the feature sets based on individual gene markers.

We also repeated the entire experiments using LDA (linear discriminant analysis), instead of logistic regression, for building the classifiers. The results are shown in Fig. 5B, where we can see similar trends as in Fig. 5A. The proposed classification method yielded the highest classification accuracy in both within-dataset experiments,

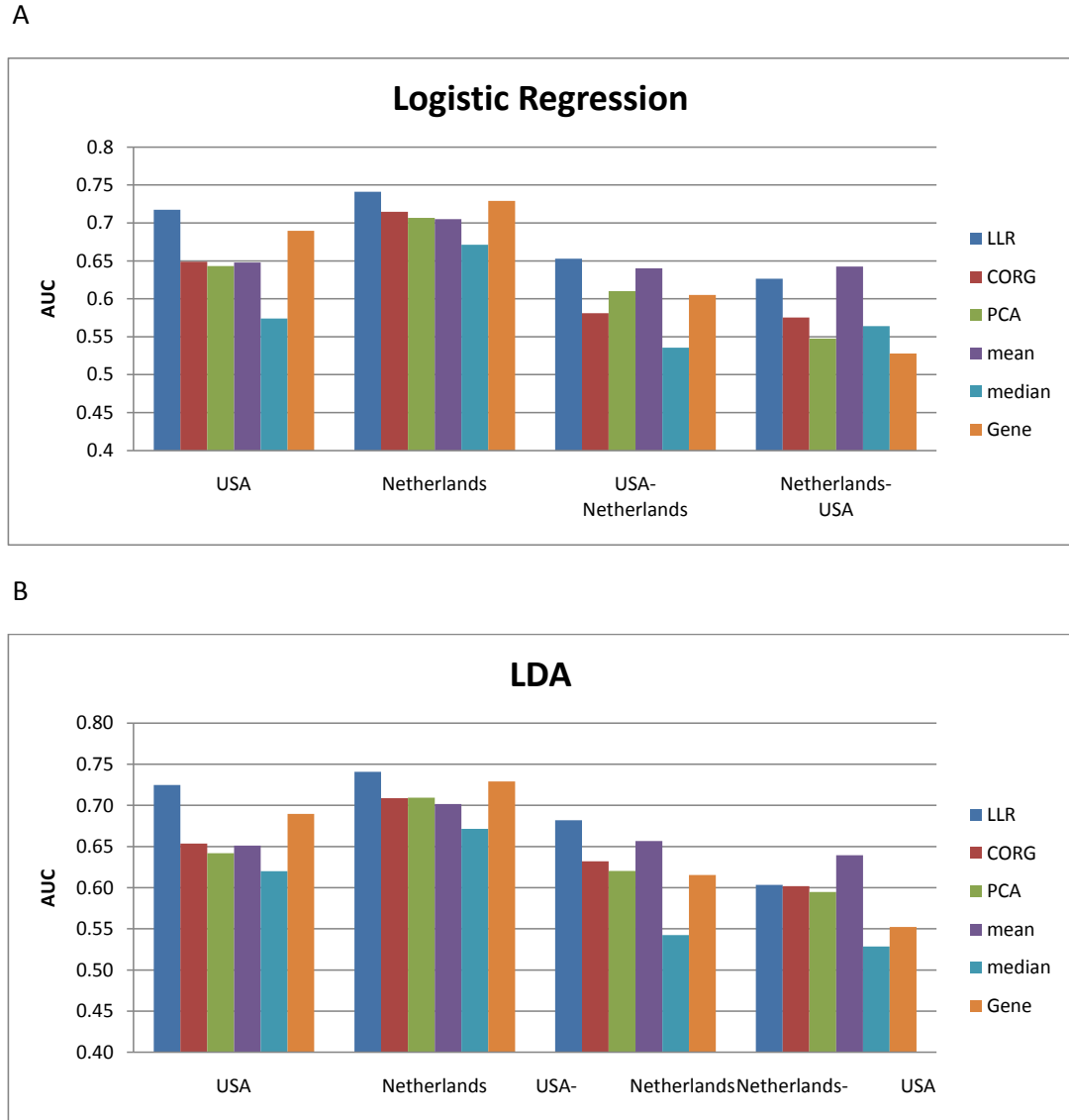


Fig. 5. Performance of different classification methods.

The bar charts show the average AUCs for different classification methods. Five pathway-based methods that use distinct pathway activity inference schemes (LLR, CORG, PCA, mean, and median) and a gene-based method were compared. (A) Classifiers were constructed based on logistic regression. Results of within-dataset experiments based on the USA and Netherlands datasets are shown in the two charts on the left. The two charts on the right show the results of the cross-dataset experiments. (B) The performance of different classification methods based on LDA (linear discriminant analysis).

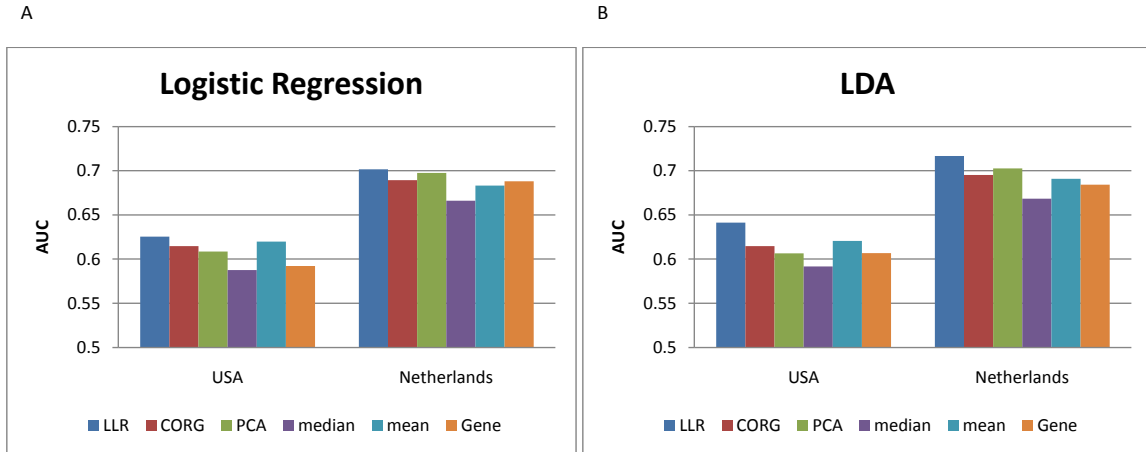


Fig. 6. Performance of different classification methods.

The bar charts show the average AUCs of within-dataset experiments for five pathway-based methods (LLR, CORG, PCA, mean, and median) and a gene-based method. In these experiments, the top 50 pathways have been reselected in every experiment using the designated training set. (A) Classification results based on logistic regression. (B) Classification results based on LDA (linear discriminant analysis).

and it also outperformed other methods in cross-dataset experiments, with the only exception of the mean-based inference method in one of the experiments.

Finally, in order to analyze the overall effect of preselecting the initial marker set, we carried out another set of within-dataset experiments, where the initial markers were reselected in every experiment using only the designated training data. The classification results are shown in Fig. 6A and 6B for logistic regression and LDA, respectively. As we can see from these figures, the preliminary marker selection step has important influence on the overall classification results, where the sensitivity of the selection method may adversely affect the performance of the resulting classifiers. However, as we can see from Fig. 6, the relative performance between different classification methods showed similar tendency as in the previous set of experiments (see Fig. 5), and the proposed method consistently outperformed the other methods in all experiments.

3. Proposed Method Leads to Robust Classifiers that Yield Symmetric Results for Dataset Inversion

Ultimately, we want to construct a robust classifier that yields accurate and consistent classification results on independent gene expression datasets. Given two independent datasets of similar size, where one dataset is used for training the classifier and the other dataset is used for evaluation, a robust classification scheme would show consistent classification performance if the training set were interchanged with the test set. However, the USA breast cancer dataset [11] and the Netherlands dataset [10] had been obtained from different microarray platforms and also preprocessed using different methods, which makes it practically difficult to evaluate the robustness of the proposed classification method by training the classifier based on one of the datasets and evaluating its performance on the other dataset. For this reason, we performed the following two-fold cross-validation experiments to assess the robustness of the proposed approach. First, we randomly divided a given dataset into two subsets of equal size. One of the subsets was used to build an actual classifier based on LDA with a classification threshold of $\lambda_{th} = 0.5$. The classifier was then used to classify the samples in the other subset and the classification error rate was computed. Next, we interchanged the training set and the test set and repeated the previous experiment. In order to find out whether we can obtain consistent classification performance after interchanging the training and test sets, we computed the absolute difference between the two classification error rates. We repeated this experiment for 250 random partitions of each breast cancer dataset, and estimated the distribution of the absolute error difference. For comparison, we carried out the above experiments using the proposed pathway activity inference scheme as well as the CORG-based scheme [26]. The proposed classification scheme resulted in a relatively small average error dif-

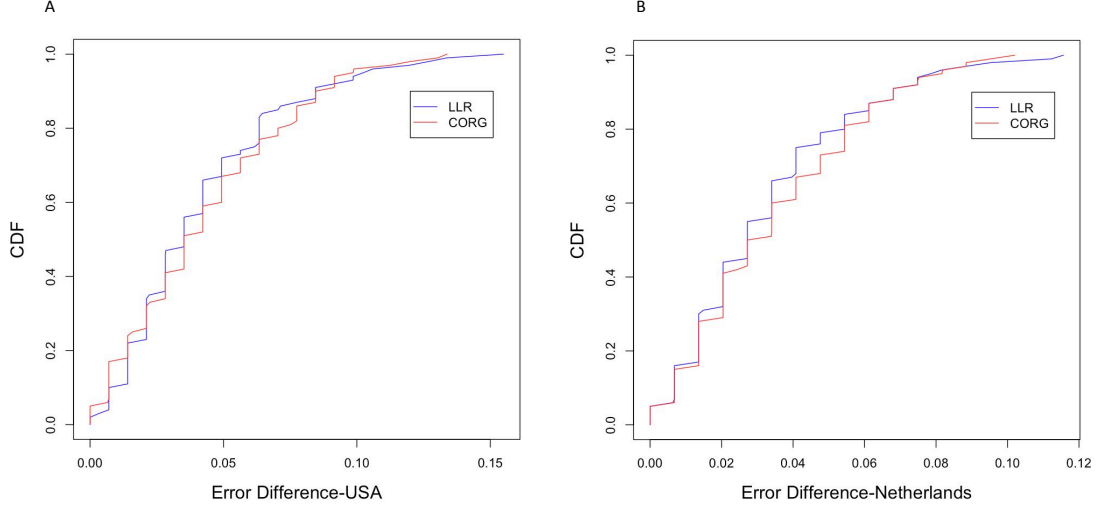


Fig. 7. Robustness of the proposed classification scheme.

To assess the robustness of the proposed classification scheme, two-fold cross-validation experiments have been performed, where we measured the change in classification error after interchanging the training and test sets. (A) Cumulative distribution of the error difference for the USA dataset. (B) Cumulative distribution of the error difference for the Netherlands dataset.

ference of 0.0414 on the USA dataset, and 0.0324 on the Netherlands dataset. The CORG-based classification scheme yielded a slightly higher error difference, whose average was 0.0429 for the USA dataset and 0.0345 for the Netherlands dataset. Figure 7 shows the cumulative distribution of the classification error difference on the two datasets for the respective methods. These results indicate that both pathway-based classification schemes can lead to the construction of robust classifiers that yield consistent results on different datasets, where the proposed scheme compares favorably to the CORG-based scheme.

C. Discussion

In this chapter, we have proposed a novel probabilistic activity inference scheme that estimates the activation level of a pathway based on the log-likelihood ratios (LLRs)

of the member genes. The proposed method can effectively address several shortcomings of the previous pathway activity inference methods, thereby improving the discriminative power of the pathway markers. For example, the methods proposed by Guo et al. [25] estimate the pathway activity by taking the mean or median of the gene expression values of the member genes. These methods cannot effectively capture the coherent gene expression patterns that may be present within a pathway. For example, suppose a member gene is positively correlated with a phenotypes of interest, while another gene in the same pathway is negatively correlated with the given phenotype. In this case, we may lose much of the discriminative information contained in the respective gene expression values if we average them out. The PCA-based inference method used in a number of studies [23, 28] can somewhat relieve this problem. In the PCA approach, the first basis vector captures the average expression pattern of the member genes, and the first principal component can estimate the presence and the strength of this pattern in a gene expression profile. However, not all the member genes may alter their expression levels under different phenotypes in a consistent manner. In fact, some genes may have expression changes that are irrelevant to the phenotypic change of our interest. To address this problem, Lee et al. [26] proposed a new pathway activity inference method that uses only a subset of member genes, called CORGs (condition-responsive genes), whose combined expression levels are highly discriminative of the phenotypes. However, the CORG-method may disregard member genes that have consistent, but not large, expression changes under different phenotypes.

The proposed LLR-based method provides an effective solution to these problems. First of all, by using the LLR of a member gene, instead of directly using its expression value, the proposed method can capture the consistent gene expression changes that are related to the phenotypic change. Moreover, since the LLR is

computed based on the difference in distribution of the gene expression values under different conditions, the direction and the amount of expression changes do not have large effects on the overall discriminative power of the pathway marker. Furthermore, the proposed method fully utilizes the available discriminative information in all the member genes, not just some of them; and it naturally weights and combines the support from each member gene in a given pathway to increase the discriminative power of the corresponding pathway marker. As we have demonstrated in this paper, the LLR-based pathway activity inference scheme significantly improves the discriminative power of the pathway markers, increases the overall classification accuracy, and finds reliable pathway markers that are more reproducible across different datasets. Therefore, the proposed method may ultimately lead to the construction of more reproducible classifiers. The two-fold cross-validation experiments, where we measured the change in classification error that resulted from interchanging the training and test sets, demonstrated the potential of the proposed scheme for building robust and reproducible classifiers.

Currently, one limitation of the pathway-based classifiers is the limited coverage of genes by known biological pathways. We believe that the classification performance of the pathway-based methods will be considerably improved once we have a more complete list of biological pathways. One possible way to overcome this problem is to identify effective pathway (or subnetwork) markers by overlaying a protein- protein interaction (PPI) network with gene expression data and searching for significantly differentially expressed regions in the given network, as proposed in [29]. In this work, we assumed that the expression values of a gene follows a Gaussian distribution. Although this has been shown to be a good approximation in our experiments, using alternative distributions that better fit the expression data may further improve the overall classification performance. For example, we may consider using gamma

distributions as proposed by Efroni et al. [33].

CHAPTER III

IDENTIFICATION OF DIAGNOSTIC SUBNETWORK MARKERS FOR CANCER IN HUMAN PROTEIN-PROTEIN INTERACTION NETWORK*

In this chapter, we propose a new method for identifying effective subnetwork markers from a PPI network by performing a *global* search for differentially expressed linear paths using dynamic programming. After finding the most discriminative linear paths, we combine overlapping paths into subnetworks through a greedy approach and use those subnetworks as diagnostic markers for classifying breast cancer metastasis. To test the effectiveness of our subnetwork markers, we perform cross validation experiments based on two independent breast cancer datasets. We compare the performance of our method with a gene-based method, a pathway-based method [34] and a previously proposed subnetwork-based method [29]. The results show that the proposed method finds reliable subnetwork markers that can accurately classify breast cancer metastasis. We also perform an enrichment analysis and show that the identified subnetwork markers are highly enriched with proteins that have common GO terms.

A. Results and Discussion

1. Identification of Subnetwork Markers

We obtained two independent breast cancer datasets from the large-scale expression studies in Wang et al. [11] (referred as the USA dataset) and van't Veer et al. [10]

*Reprinted with permission from “Identification of Diagnostic Subnetwork Markers for Cancer in Human Protein-Protein Interaction Network” by J. Su, B.J. Yoon and E.R. Dougherty, *BMC Bioinformatics* 2010, 11(Suppl 6):S8. Copyright 2010 by BMC Bioinformatics.

(referred as the Netherlands dataset). The USA dataset contains 286 samples and the Netherlands dataset contains 295 samples. Metastasis had been detected for 78 patients in the Netherlands dataset and 107 patients in the USA dataset during the five-year follow-up visits after the surgery. The PPI network has been obtained from Chuang et al. [29], which contains 57,235 interactions among 11,203 proteins. Since not all proteins have corresponding genes in the microarray platforms used by the two breast cancer studies, we used the induced network which contains 9,263 proteins and 49,054 interactions for the USA dataset, and 8,380 proteins and 31,201 interactions for the Netherlands dataset.

Our proposed method integrates the gene expression data and the PPI data by overlaying the expression value of each gene on its corresponding protein in the PPI network. The subnetwork identification algorithm consists of the following three major steps:

Step 1: Search for highly discriminative linear paths whose member genes are closely correlated to each other

To find discriminative linear paths in the large PPI network, we define a scoring scheme that incorporates both the t -test statistics scores of the member genes and the correlation coefficient between their expression values. This scoring scheme takes a weighted sum of the t -scores of the member genes within a given path. The weights depend on the correlation between the member genes and the parameter θ , where θ is introduced to control the trade off between the “discriminative power” of individual genes and the “correlation” between the member genes (see Methods). Based on the above scoring scheme, we developed an algorithm that searches for the top scoring linear paths that have length ℓ and end at node g_i .

Step 2: Combine top scoring linear paths into a subnetwork

We initialize the subnetwork using the path with the highest score. As long

as there exists a high scoring path that overlaps with the current subnetwork, we combine them and check if the discriminative power of the new subnetwork is larger than that of the previous subnetwork. If the discriminative power improves, we keep the new subnetwork. Otherwise, we keep the previous subnetwork and check the next best path. To evaluate the discriminative power of subnetworks, we applied the probabilistic pathway activity inference method proposed in [34] to infer the subnetwork activity. The discriminative power of a subnetwork is assessed by computing the t -test statistics score of the subnetwork activity.

Step 3: Update the PPI network

After identifying the discriminative subnetwork, we update the PPI network by removing the proteins in the identified subnetwork from the current PPI network. In order to find additional *non-overlapping* subnetworks, we repeat the search process from **Step 1**.

In order to control the size of the identified subnetworks, we restricted the length of the linear paths to be less than 8. For a given ℓ and for every node g_i in the network, we identified the top 20 linear paths with the highest scores, whose length is ℓ and end at the given node g_i . To construct the subnetwork marker that can be used as a diagnostic marker for breast cancer metastasis, we chose the top 100 scoring linear paths whose length are within a given range $5 \leq \ell \leq 8$. The selected linear paths were combined into a single subnetwork as described in **Step 2**. To find the best θ , we repeated the experiment for six different values $\theta = 1, 2, 4, 8, 16$ and ∞ . For every value of θ , we identified 50 subnetwork markers for each dataset using the proposed method. The statistics of the identified subnetworks for the two datasets are shown in Table I. We can see that the overlap between the subnetwork markers identified on different datasets is around 25%, which is significantly larger than the overlap reported in Chuang et al. (12.7%) [29].

Table I. Statistics of the subnetwork markers identified by the proposed method.

θ		Size		Number of genes	Number of genes in common
		mean	standard deviation		
1	USA	16.8	10.17	840	213
	Netherlands	14.62	8.69	731	
2	USA	18.22	12.3	911	233
	Netherlands	16	10.34	801	
4	USA	18	12.8	901	202
	Netherlands	17.28	11.4	864	
8	USA	20.7	13.38	1035	252
	Netherlands	19.52	12.57	976	
16	USA	20.2	11.13	1010	201
	Netherlands	16.64	10.89	832	
∞	USA	22.32	14.86	1116	266
	Netherlands	21.92	10.67	1096	

For each θ , we show the mean and standard deviation of the subnetwork size as well as the total number of genes covered by the identified subnetworks. We also show the number of genes shared by the subnetworks identified using the respective breast cancer datasets.

2. The Identified Subnetworks are Enriched with Proteins in Common GO Terms

We identified 50 discriminative subnetworks using the proposed method for both the USA dataset and the Netherlands dataset ($\theta = 8$). The identified subnetworks consist of 1035 and 976 genes, respectively. Next, we analyzed the identified subnetworks using FuncAssociate [35], which is a web application designed for characterizing large collections of genes and proteins. It performs a Fisher’s Exact Test (FET) analysis to identify Gene Ontology (GO) [24] attributes that are shared by a fraction of the entries in a given set of genes or proteins. At a significance threshold of 0.01, 78% and 84% of the subnetworks that were respectively identified using the USA dataset and the Netherlands dataset were enriched with proteins that share common GO terms. These GO terms generally correspond to cell growth and death, cell proliferation and replication, cell and tissue remodeling, circulation and coagulation, or metabolism. Examples of the identified subnetworks are shown in Figure 8, where we can see that the proposed method is capable of finding subnetwork markers that also include genes that are oppositely regulated. The enrichment analysis results of the sample subnetworks obtained using FuncAssociate are shown in Table II.

3. Subnetwork Markers Identified by the Proposed Method are More Discriminative and Reproducible

We first evaluated the subnetwork markers identified using the proposed method. For a given θ , we identified the subnetwork markers based on one dataset and estimated their discriminative power on the same dataset. The discriminative power of the subnetwork marker was estimated as the absolute t -test statistics score of the subnetwork activity. Subnetwork markers were then sorted in the decreasing order of t -score. Next, to show the reproducibility of our subnetwork markers, we identified

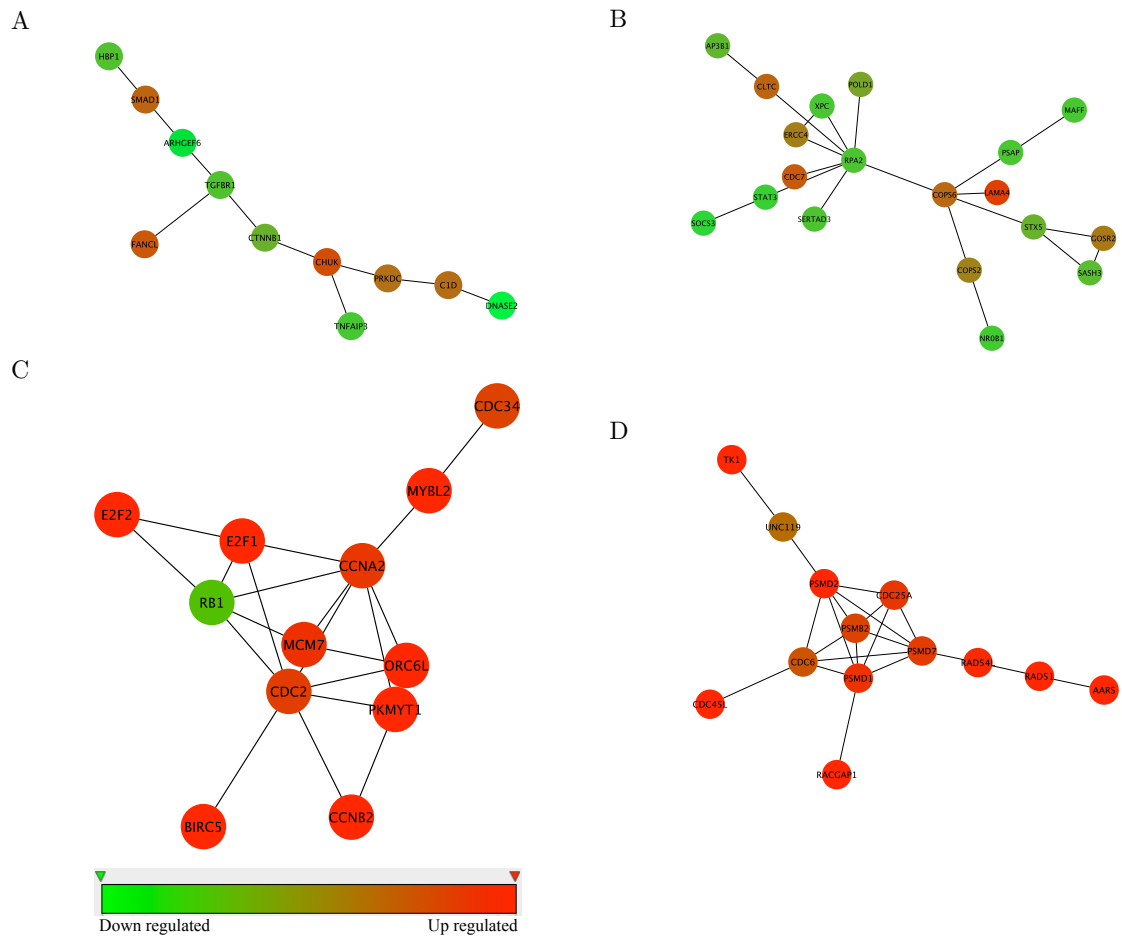


Fig. 8. Sample subnetworks identified using the proposed method.

(A),(B) are examples of subnetworks identified using the USA dataset. (C),(D) are examples of subnetworks identified using the Netherlands dataset. Red (green) implies that the gene is upregulated (downregulated) in breast cancer samples with metastasis.

Table II. Enrichment analysis results for the sample subnetworks shown in Figure 8.

Subnetwork	Attribute ID	<i>P</i> -value	Attribute name
A	GO:0045165	0.024	cell fate commitment
	GO:0012501	0.001	programmed cell death
	GO:0008219	0.006	cell death
	GO:0016265	0.006	death
	GO:0006915	0.017	apoptosis
B	GO:0000718	< 0.001	nucleotide-excision repair, DNA damage removal
	GO:0006308	0.046	DNA catabolic process
	GO:0043566	0.040	structure-specific DNA binding
C	GO:0051318	0.039	G1 phase
	GO:0022403	< 0.001	cell cycle phase
	GO:0005654	< 0.001	nucleoplasm
	GO:0000280	0.009	nuclear division
	GO:0007067	0.009	mitosis
	GO:0048285	0.009	organelle fission
	GO:0051301	0.001	cell division
	GO:0022402	< 0.001	cell cycle process
	GO:0007049	0.000	cell cycle
	GO:0051726	0.008	regulation of cell cycle
	GO:0044428	0.001	nuclear part
D	GO:0005838	< 0.001	proteasome regulatory particle
	GO:0000076	0.016	DNA replication checkpoint
	GO:0032297	0.016	negative regulation of DNA replication initiation

Table II. Continued.

Subnetwork	Attribute ID	<i>P</i> -value	Attribute name
	GO:0030174	0.030	regulation of DNA replication initiation
	GO:0031145	< 0.001	anaphase-promoting complex-dependent proteasomal ubiquitin-dependent protein catabolic process
	GO:0051436	< 0.001	negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle
	GO:0051352	< 0.001	negative regulation of ligase activity
	GO:0051437	< 0.001	positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle
	GO:0051444	< 0.001	negative regulation of ubiquitin-protein ligase activity
	GO:0051443	0.001	positive regulation of ubiquitin-protein ligase activity
	GO:0051439	0.001	regulation of ubiquitin-protein ligase activity during mitotic cell cycle
	GO:0051351	0.001	positive regulation of ligase activity
	GO:0051438	0.002	regulation of ubiquitin-protein ligase activity
	GO:0051340	0.002	regulation of ligase activity
	GO:0010498	0.007	proteasomal protein catabolic process
	GO:0043161	0.007	proteasomal ubiquitin-dependent protein catabolic process
	GO:0022402	< 0.001	cell cycle process

the top 50 markers based on one dataset and evaluated their discriminative power on the other dataset. Again, subnetwork markers were sorted according to their discriminative power. Figure 9 shows the discriminative power of subnetwork markers identified using six different values of θ , where the x -axis corresponds to the top K markers being considered, and the y -axis shows the mean absolute t -score of the top K markers ($K = 10, 20, 30, 40, 50$). Figure 9A and Figure 9B show the results obtained from the USA dataset and the Netherlands dataset, respectively. Figure 9C shows the discriminative power of the subnetwork markers selected based on the Netherlands dataset and evaluated using the USA dataset. Figure 9D shows the discriminative power of the markers selected based on the USA dataset and evaluated using the Netherlands dataset. As we can see from these results, the discriminative power of the identified subnetwork markers is not very sensitive to the choice of θ . To further compare the identified subnetwork markers with other markers, we used $\theta = 8$ which showed good performance in average.

Next, we compared the identified subnetwork markers with gene markers, pathways markers [34] and the subnetwork markers identified by Chuang et al. [29]. For gene markers, we selected the top 50 genes based on the absolute t -score among all genes covered by the 50 identified subnetworks. For pathway markers, we selected the top 50 pathways among the 639 pathways in the C2 curated gene sets in MsigDB (Molecular Signatures Database) [21]. We also obtained the subnetworks identified by Chuang et al. [29] from the Cell Circuits database [36] (149 discriminative subnetworks for the Netherlands dataset and 243 subnetworks for the USA dataset). We chose the top 50 subnetworks out of 149 subnetworks based on the Netherlands dataset and the top 50 subnetworks out of 243 subnetworks based on the USA dataset. The pathways and subnetworks were ranked using the scheme proposed by Tian et al. [22], based on the average absolute t -test statistics score of all the member genes. For subnetwork

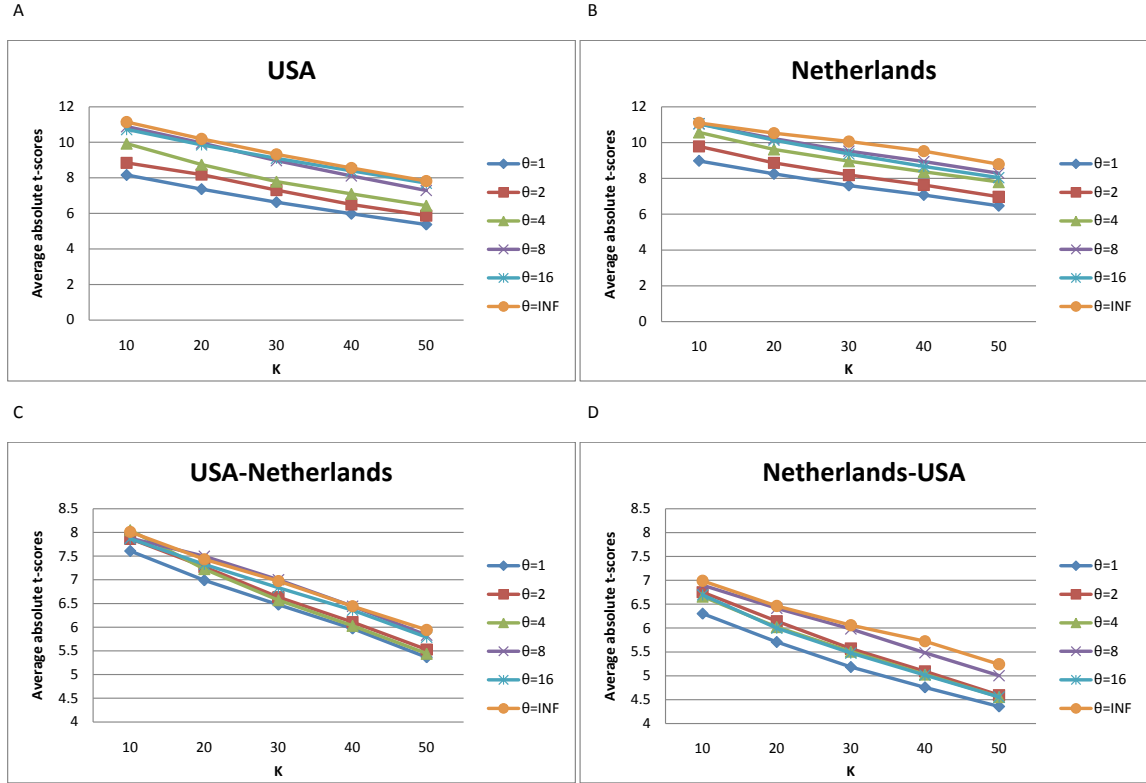


Fig. 9. Discriminative power of the subnetwork markers identified by the proposed method using different θ .

We computed the mean absolute t -score of the top $K = 10, 20, 30, 40, 50$ subnetwork markers for different values of θ (shown in different colors). (A), (B): Markers were identified using a particular dataset and tested on the same dataset. (C), (D): Markers were identified using the first dataset and evaluated on the second dataset.

markers identified by Chuang et al., we computed the t -scores of their member genes using the original expression values. For pathway markers, t -scores of the member genes were computed using their log-likelihood ratios as in [34] (see Methods).

To assess the discriminative power of the subnetwork markers identified using the proposed method, their activity score was inferred using the probabilistic inference method proposed in [34]. For subnetwork markers identified by Chuang et al., we inferred their activity score using the mean expression value of the member genes as reported in their paper [29].

The discriminative power of these different markers are shown in Figure 10. As we can see in Figure 10, subnetwork markers identified by our method are more discriminative compared to other markers. Moreover, it can be seen that they also retain higher discriminative power across different datasets.

4. Subnetwork Markers Identified by the Proposed Method Improves Classification Performance

To evaluate the performance of the classifiers that are constructed using the subnetwork markers identified by the proposed method, we performed the following within-dataset and cross-dataset cross-validation experiments.

In the within-dataset experiments, the top 50 subnetwork markers identified using one of the two breast cancer datasets were used to build the classifier. The dataset was divided into ten folds of equal size, one of them was withheld as the “test set” and the remaining nine were used for training the classifier. In the training set, six folds (referred as the “marker ranking set”) were used to rank the subnetwork markers according to their discriminative power and to build the classifier using logistic regression. The other three folds (referred as the “feature selection set”) were used for feature selection. We started with the top ranked subnetwork marker and enlarged

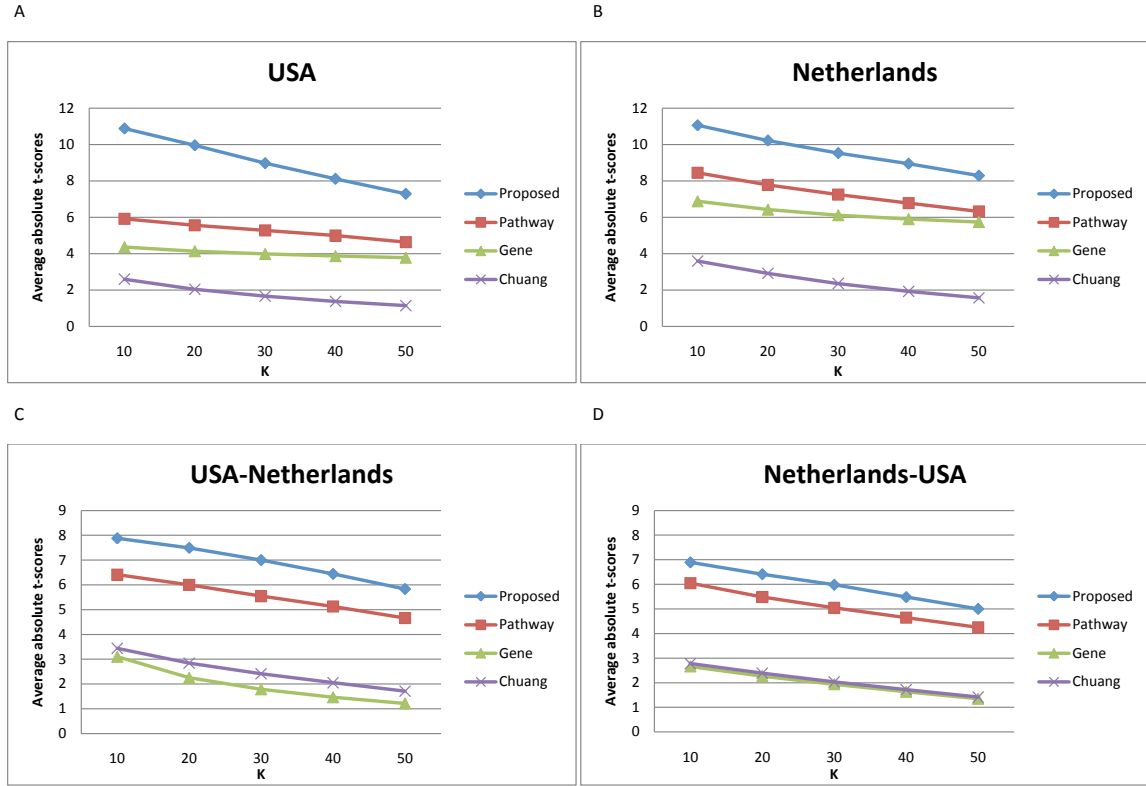


Fig. 10. Discriminative power of different types of markers.

We evaluated the discriminative power of the subnetwork markers identified using the proposed method, and compared them with gene markers, pathway markers[34], and the subnetwork markers identified by Chuang et al.[29]. Mean absolute t -score is shown for the top $K = 10, 20, 30, 40, 50$ markers. (A), (B): Markers were identified using a particular dataset and tested on the same dataset. (C), (D): Markers were identified using the first dataset and evaluated based on the second dataset.

the feature set by adding features sequentially. Every time we included a new subnetwork marker into the feature set, a new classifier was built using the marker ranking set and it was tested on the feature selection set. For all the samples in the feature selection set, the classifier can compute the posterior probabilities of the class label (metastasis versus metastasis-free), based on which we can estimate the AUC (Area Under ROC Curve) [32]. The AUC metric provides a useful statistical summary of the classification performance over the entire range of sensitivity and specificity. We retained the new subnetwork marker if the AUC (estimated on the feature selection set) increased; otherwise, we discarded the subnetwork marker and continued to test the remaining ones. The above experiment was repeated 500 times based on 50 random ten-fold splits. The average AUC was reported as the classification performance measure.

To evaluate the reproducibility of the subnetwork markers, we performed the following cross-dataset experiments. We first identified the top 50 subnetwork markers based on one dataset and performed cross-validation experiments on the other dataset, following a similar procedure that was used in the previously described within-dataset experiments.

For comparison, we also performed similar within-dataset and cross-dataset experiments using gene markers, pathway markers and the subnetwork markers identified by Chuang et al., respectively. For each method, we limited the feature set to the top 50 markers for each dataset. Figure 11 shows the classification performance based on the subnetwork markers identified by the proposed method for different values of θ . We found that the AUC for both within-dataset and cross-dataset experiments first increases with increasing θ and starts to drop after certain point. At $\theta = 8$, the AUC values for both cross-dataset experiments are relatively larger than those at other values of θ . Also, the AUC values for both within-dataset experiments at $\theta = 8$

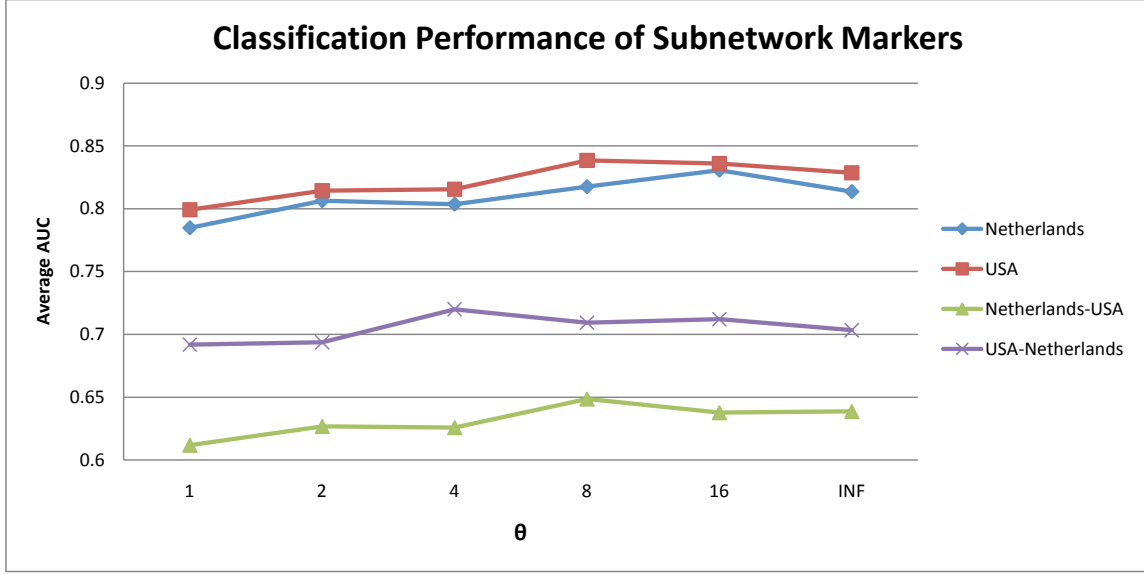


Fig. 11. Classification performance of the identified subnetwork markers for different θ .

The line plots show the average AUC for classifiers based on subnetwork markers identified using $\theta = 1, 2, 4, 8, 16, \infty$. The legends USA, Netherlands denote the results of within-dataset experiments based on the USA dataset and the Netherlands dataset, respectively. The legends USA-Netherlands, Netherlands-USA denote the results of cross-dataset experiments where markers were identified based on the first dataset and tested based on the second dataset.

compare favorably with those at different θ , which implies that the trade off between maximizing the discriminative power and increasing the correlations of the member genes is well balanced.

To compare the classification performance of the identified subnetwork markers with other types of markers, we set $\theta = 8$. Based on this setting, we compared our subnetwork markers with gene markers, pathway markers and the subnetwork markers from Chuang et al. using the experimental designs described above. Figure 12 summarizes the classification performance of the proposed approach, in comparison with the other methods. The two bar charts on the left of Figure 12 show the AUC of the within-dataset experiments. As shown in Figure 12, classifiers based on the subnetwork markers identified by the proposed method perform significantly

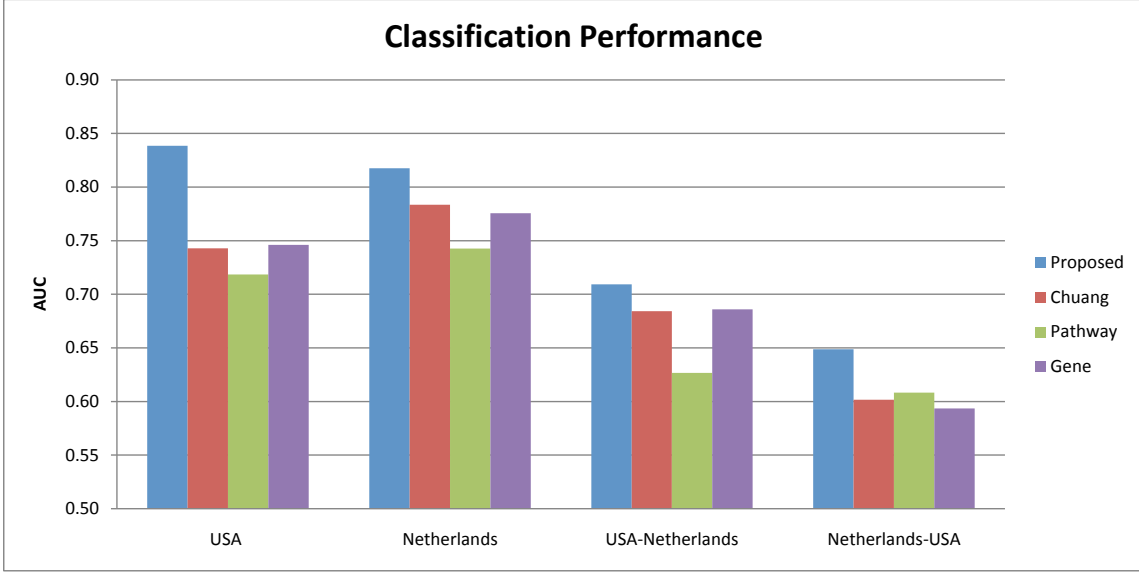


Fig. 12. Classification performance of different types of markers.

The bar charts show the average AUC of different classifiers that use subnetwork markers identified by the proposed method, gene markers, pathway markers, and subnetwork markers found by Chuang et al.'s method. Results of the within-dataset experiments based on the USA and Netherlands dataset are shown in the two bar charts on the left. The two bar charts on the right show the results of the cross-dataset experiments, where markers were identified based on the first dataset and tested based on the second dataset.

better than the classifiers based on other types of markers. The results of the cross-dataset experiments are shown in the two bar charts on the right of Figure 12. Again, we can see that the classifiers built on the subnetwork markers predicted by our method significantly outperform those based on other markers. This indicates that the predicted subnetwork markers are more reproducible compared to other markers.

Figure 13 shows the classification error of the classifiers built using different types of markers at different TPR (true positive rate). As shown in Figure 13, the error curve that corresponds to the proposed markers always lies below others, which implies that classifiers built on our subnetwork markers yield a lower error rate at any fixed sensitivity level.

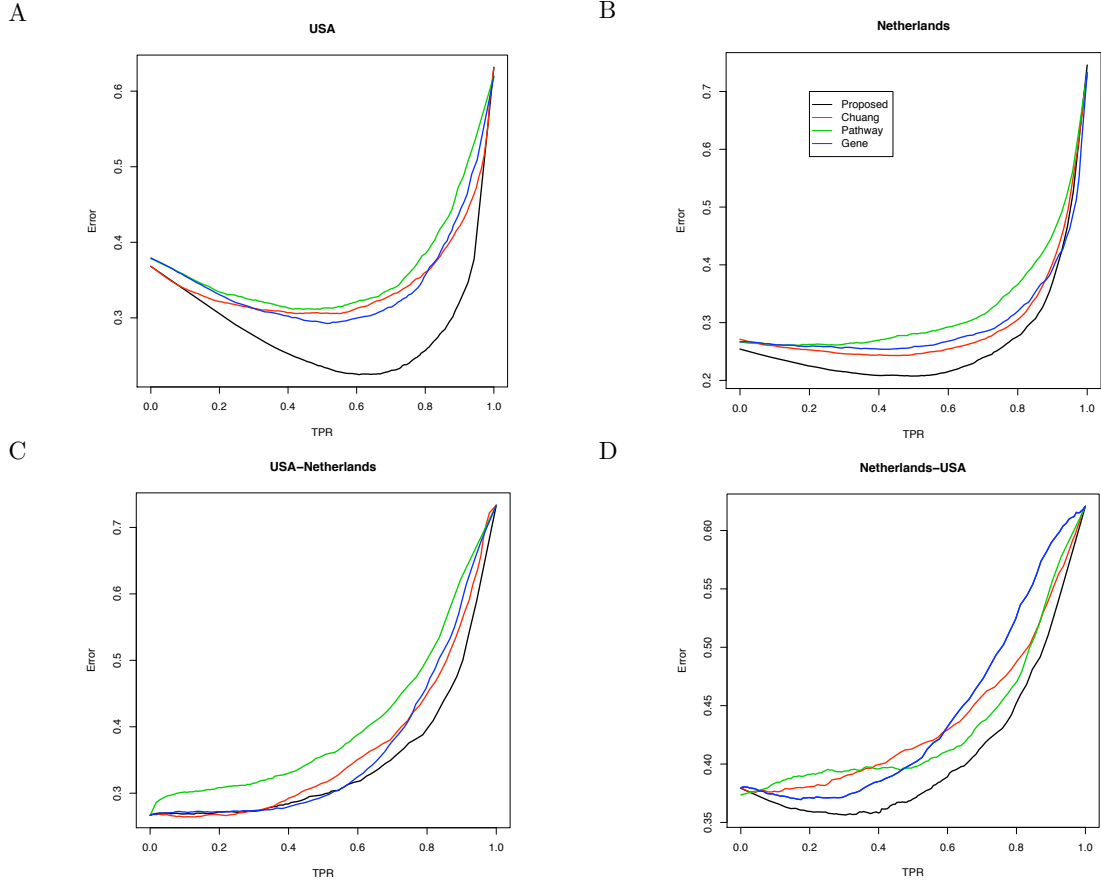


Fig. 13. Classification error at different TPR (true positive rate) for different types of markers.

(A), (B) show the results of the within-dataset experiments based on the USA dataset and the Netherlands dataset, respectively. (C), (D) show the results of the cross-dataset experiments, where markers were identified using the first dataset and tested based on the second dataset.

B. Conclusions

In this paper, we proposed a new method for identifying effective subnetwork markers in a protein-protein interaction (PPI) network. As shown throughout this paper, integrating the PPI network with microarray data can overcome some of the shortcomings of the gene-based and pathway-based methods. First of all, using a genome-scale PPI network provides a better coverage of the genes in the microarray studies compared to using known pathways obtained from public databases. Second, the network topology provides prior information about the relationship between proteins, hence the genes that code for these proteins. Subnetworks identified by integrating the network structure and the gene expression data can cluster proteins (or genes) that are functionally related to each other. By aggregating the expression values of the member genes, subnetwork markers can avoid selecting single gene markers with redundant information. Furthermore, the discriminative subnetworks identified by the proposed method can also provide us with important clues about the biological mechanisms that lead to different disease phenotypes.

The proposed method finds top scoring linear paths using dynamic programming and combines them into a subnetwork by greedily optimizing the discriminative power of the resulting subnetwork marker. We developed a scoring scheme that is used by the search algorithm to find linear paths that consist of discriminative genes that are highly correlated to each other. The proposed algorithm allows us to control the trade off between maximizing the discriminative power of the member genes within a given linear path and increasing the correlation between the member genes, by choosing the appropriate value for θ . As the subnetwork markers are constructed based on the top scoring linear paths, instead of single genes, the proposed method is expected to yield more robust subnetwork markers. Another important advantage of our method

is that it can find non-overlapping subnetwork markers. This can reduce the overall redundancy among the identified markers. In this paper, the activity of the identified subnetwork markers were inferred using the probabilistic activity inference scheme proposed in [34]. This allows us to find better subnetwork markers, since it can assess their discriminative power more effectively.

As shown in this paper, the identified subnetwork markers consist of proteins that share common GO terms. The classifiers based on the subnetwork markers identified using the proposed method were shown to achieve higher classification accuracy in both within-dataset and cross-dataset experiments compared to classifiers based on other markers. These results suggest that the method proposed in this paper can find effective subnetwork markers that can more accurately classify breast cancer metastasis and are more reproducible across independent datasets.

C. Methods

1. Overview

Given a large PPI network, we want to find subnetwork markers whose activity is highly indicative of the disease state of interest. For this purpose, we first need a method for inferring the activity of a given subnetwork and evaluating its discriminative power. There exist different ways for computing the activity score of a given group of genes [34]. Recently, we proposed a probabilistic pathway activity inference scheme, which was shown to outperform many other existing methods. Thus, we adopt this activity inference scheme for finding subnetwork markers whose activity scores are highly discriminative of the disease states. However, finding the subnetwork markers with maximum discriminative power in a PPI network based on the selected inference method is computationally infeasible. For this reason, we propose

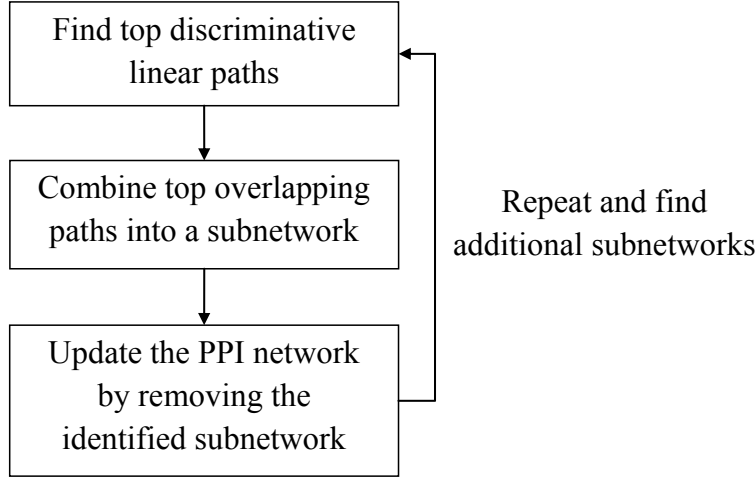


Fig. 14. Illustration of the proposed method.

an algorithm for identifying effective subnetwork markers which is motivated by a simple scheme proposed in Tian et al. [22]. This scheme scores a pathway marker by computing the average absolute t -score of its member genes. It has been shown to be effective in evaluating the discriminative power of pathway markers in [34]. Since our goal is to find groups of genes that display coordinated expression patterns, we modified Tian et al.'s scoring scheme to incorporate the correlation between the genes within a given pathway. This new method scores a given pathway by taking the weighted sum of the absolute t -scores of its member genes, where the weights are computed using the correlation coefficients between the member genes.

The general outline of the proposed algorithm is as follows. Based on the above scoring scheme, we first search for differentially expressed linear paths in the PPI network. Then, the top paths that overlap with each other are greedily combined into a subnetwork by maximizing the discriminative power of the resulting subnetwork, evaluated by the method proposed in [34]. The identified subnetwork is removed from the PPI network, and the above process is repeated to find multiple non-overlapping subnetwork markers. The overall scheme is illustrated in Fig. reffigure.

2. Probabilistic Inference of Subnetwork Activity

Here we provide a brief review of the probabilistic activity inference method proposed in [34]. Suppose we have a subnetwork \mathcal{G}_s that consists of n proteins which correspond to n different genes $\{g_1, g_2, \dots, g_n\}$. Assume that the expression level x_i of a gene g_i follows the distribution $f_i^k(x_i)$ under phenotype $k = 1, 2$. The log-likelihood ratio (LLR) [34] between the two phenotypes is computed as follows

$$\alpha(x_i) = \log(f_i^1(x_i)/f_i^2(x_i)).$$

In order to estimate the conditional probability density function $f_i^k(x_i)$, we assume that the gene expression level of gene g_i under phenotype k follows a Gaussian distribution with mean μ_i^k and standard deviation σ_i^k . The parameters are empirically estimated using the samples with phenotype k . Given the log-likelihood ratio of each gene, the subnetwork activity $A_{\mathcal{G}_s}$ is defined as the sum of the log-likelihood ratios of the member genes $A_{\mathcal{G}_s} = \sum_{i=1}^n \alpha(x_i)$.

3. Evaluating the Discriminative Power of Linear Paths in the PPI Network

A linear path $\lambda = \{g_1, g_2, \dots, g_n\}$ in a given PPI network \mathcal{G} is defined as a group of genes, where the proteins that correspond to g_i and g_{i+1} are connected for $i = 1, \dots, n-1$. To evaluate the discriminative power of a linear path, we first evaluate the discriminative power of each gene g_i by computing the t -test statistics score of the log-likelihood ratio $\alpha(x_i)$, denoted as $t_\alpha(g_i)$. Then, we compute the Pearson product-moment correlation coefficient to measure the correlation between the log-likelihood

ratios of $\forall g_i, g_j \in \lambda$. The correlation matrix is given by

$$\Sigma(\lambda) = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & 1 & \cdots & \rho_{2n} \\ \vdots & \vdots & & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & 1 \end{bmatrix}$$

where $\rho_{ij}, i \neq j$ is the correlation coefficient between the log-likelihood ratios of g_i and g_j . The score of the pathway λ is defined as following

$$S(\lambda) = \frac{1}{n^2} [t_\alpha(g_1), t_\alpha(g_2), \dots, t_\alpha(g_n)] \cdot \Sigma'(\lambda)$$

where $\Sigma'(\lambda) = \frac{1}{1+\theta} [(\Sigma(\lambda) - I) + \theta \cdot I]$ and I is the identity matrix. We use a normalization factor of $\frac{1}{n^2}$ to ensure that the overall score does not depend on the length of the path. We use θ to control the trade off between maximizing the discriminative power of the genes within the identified path and increasing the correlation between its member genes. When $\theta = 0$, the weight for the t -score of a given gene g_i is determined by the average correlation between the log-likelihood ratios of g_i and g_j , where $j \neq i$. As θ increases, we give more weight on the discriminative power of individual genes than the correlation between member genes. Especially, when $\theta \rightarrow \infty$, we get $\Sigma'(\lambda) = I$. In this case, the pathway score $S(\lambda)$ is simply the average t -score of the member genes in λ , and the proposed subnetwork marker identification method reduces to its preliminary version proposed in [37]. The above scoring scheme is used for finding the top linear paths in the network \mathcal{G} as we describe in the following section.

4. Searching for Discriminative Linear Paths

Let $\mathcal{G} = (E, V)$ denote the PPI network, where V is the set of nodes (i.e., proteins), E is the set of edges (i.e., protein interactions). Suppose there are N proteins in \mathcal{G} .

Then we can represent E as an N -dimensional binary matrix. For any protein pair (v_a, v_b) , where $v_a, v_b \in V$, we let $E[v_a, v_b] = 1$, if v_a, v_b are connected; $E[v_a, v_b] = 0$, otherwise. Based on the scoring scheme defined in the previous section, we search for top discriminative linear paths using dynamic programming. We define $\lambda(v_i, \ell)$ as the optimal linear path among all linear paths that have length ℓ and end at v_i . The score of this optimal path is defined as

$$s(v_i, \ell) = t_\alpha[\lambda(v_i, \ell)] \cdot \Sigma'[\lambda(v_i, \ell)].$$

Here, only paths with length $\ell \leq L$ are considered. The algorithm is defined as follows.

(i) **Initialization:** $\ell = 1, \forall v_i \in V$,

$$s(v_i, \ell) = |t_\alpha(v_i)|.$$

(ii) **Iteration:**

for $l = 2$ to L ,

for $\forall v_i \in V$,

$$s(v_i, \ell) = \max_{v_j} \{t(\lambda(v_j, \ell), v_i) \cdot \Sigma'(\lambda(v_j, \ell), v_i) + \log(E[v_i, v_j])\},$$

$$v_j^* = \arg \max_{v_j} \{t(\lambda(v_j, \ell), v_i) \cdot \Sigma'(\lambda(v_j, \ell), v_i) + \log(E[v_i, v_j])\},$$

if $s(v_i, \ell) > 0$, then

$$\lambda(v_i, \ell) = \lambda(v_j^*, \ell - 1) \cup \{v_i\}.$$

end

end

(iii) **Termination:**

for $\forall v_i \in V, 1 \leq \ell \leq L$,

$$S(\lambda(v_i, \ell)) = s(v_i, \ell)/\ell^2. \quad (3.1)$$

Although the above algorithm finds only the top path for every (v_i, ℓ) , we can easily modify it to find the top M discriminative paths. Increasing M allows us to find better linear paths with higher discriminative power, but it will also increase the computational complexity of the algorithm.

5. Combining Top Overlapping Paths into a Subnetwork

Based on (3.1), we choose the m top scoring paths $\Lambda = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$ whose length is within a given range $[L_{\min}, L_{\max}]$. Next, the paths in Λ are combined into a subnetwork \mathcal{G}_s so that its discriminative power $\mathcal{R}(\mathcal{G}_s)$ is locally optimized. This process is carried out as follows:

- (i) $\mathcal{G}_s \leftarrow \lambda_i, \mathcal{G}_{temp} \leftarrow \mathcal{G}_s, i = 1$.
- (ii) $i = i + 1$; If $\lambda_i \cap \mathcal{G}_s \neq \emptyset$, $\mathcal{G}_{temp} \leftarrow \mathcal{G}_{temp} \cup \lambda_i$.
- (iii) If $\mathcal{R}(\mathcal{G}_{temp}) > (1 + \epsilon)\mathcal{R}(\mathcal{G}_s)$, $\mathcal{G}_s \leftarrow \mathcal{G}_{temp}$; else $\mathcal{G}_{temp} \leftarrow \mathcal{G}_s$.
- (iv) Go to (ii) if $i < m$; otherwise, terminate.

Here ϵ is set as 0.01 to avoid over-fitting to the expression data. We used the activity inference method in [34] to compute the actual activity score of \mathcal{G}_s . Then, $\mathcal{R}(\mathcal{G}_s)$ is computed as the t -test statistics of the subnetwork activity score.

After obtaining a subnetwork \mathcal{G}_s , we removed it from the network \mathcal{G} by setting $E[v_s, v_i] = E[v_i, v_s] = 0, \forall v_s \in \mathcal{G}_s, v_i \in \mathcal{G}$. Then, the whole process was repeated using the updated network to find additional subnetwork markers.

CHAPTER IV

GENE CLUSTERING IN THE PPI NETWORK BASED ON A MESSAGE PASSING ALGORITHM TOWARDS ACCURATE DISEASE CLASSIFICATION

In Chapter III, we proposed a subnetwork identification algorithm based on dynamic programming. The proposed algorithm consists of three steps. First, the algorithm performs a global search for differentially expressed linear paths whose member genes are closely correlated with each other using dynamic programming. Secondly, it combines top scoring linear paths into a subnetwork. Thirdly, it updates the PPI network by removing the identified subnetwork. This algorithm finds non-overlapping subnetwork markers by repeating the above procedures. It has been proved that the identified subnetwork markers are more accurate and reliable compared to single gene markers, pathway markers and subnetwork markers identified by an existing method.

However, the proposed algorithm also has some possible defects for the following reasons. One main problem is that this algorithm finds non-overlapping subnetworks by removing the identified ones and repeating the whole procedure, which actually assumes that the identified subnetworks are all better than the following ones and therefore they have no need to join the following competitions. This may not be true because the previous subnetworks were identified by combining top scoring linear paths via a greedy approach. The heuristic used here

To address this problem, we will apply a clustering algorithm, named "Affinity Propagation", to identify gene clusters based on the PPI network. Affinity propagation takes the real-valued measures of similarity between data points as input. It considers all data points as potential exemplars simultaneously and makes the decision by exchanging two types of messages between data points which represents two different kinds of competitions. To apply affinity propagation to clustering genes, we

first define the similarity between genes that are connected within a given number of steps, then the preferences are set to be a common value for all the genes in the network. The identified clusters with highest discriminative powers are used to classify the samples in two independent breast cancer datasets. The result shows that the identified gene cluster achieve better classification performance compared to the subnetworks identified using the method proposed in Chapter III. We also compared the identified clusters with the previously identified subnetworks. The result shows that affinity propagation is more likely to find the right combinations of genes which yield better classification results based on the PPI network.

A. Methods

1. Affinity Propagation

Affinity propagation [38] is a clustering algorithm that exchanges real-valued messages between data points recursively until a good set of exemplars and corresponding clusters emerges. The input of affinity propagation is the similarities between pairs of data points. For data points with index i and k , the similarity $s(i, k)$ measures how well data point k is suited to be the exemplar for data point i . In stead of specifying the number of clusters prior to the algorithm, affinity propagation requires a real number input $s(k, k)$ for each data point which is referred to as "preference". Data points with larger preferences are more likely to be chosen as exemplars. There are two types of messages passing between data points, called "responsibility" and "availability". The responsibility $r(i, k)$ is sent from data point i to data point k , which measures how well the data point k is suitable to be exemplar for data point i considering competition from the other potential exemplars(Fig 1). The availability $a(i, k)$ is sent from data point i to data point k , which measures how appropriate it

would be for data point i to choose data point k as its exemplar, taking into account the support information from other data points who should choose data point k . The definition of $r(i, k)$ and $a(i, k)$ are as follows:

$$r(i, k) \leftarrow s(i, k) - \max_{k' \text{ s.t. } k' \neq k} \{a(i, k') + s(i, k')\} \quad (4.1)$$

$$a(i, k) \leftarrow \min\{0, r(k, k) + \sum_{i' \text{ s.t. } i' \notin \{i, k\}} \max\{0, r(i', k)\}\} \quad (4.2)$$

The "self-availability" $a(k, k)$ is updated in a different way,

$$a(k, k) \leftarrow \sum_{i' \text{ s.t. } i' \neq k} \max\{0, r(i', k)\} \quad (4.3)$$

The availability $a(i, k)$ is initially set to be 0, the two types of messages are updated recursively in the later iterations. Decisions can be made at any point during affinity propagation. The data point k that maximize $a(i, k) + r(i, k)$ is chosen to be the exemplar for point i if $k \neq i$. If $k = i$, it indicates that i is an exemplar. The algorithm converges if the set of exemplars doesn't not change for a given number of iterations.

2. Computing the Similarities between Genes

To apply Affinity Propagation to cluster genes based on the PPI network, we first need to compute the similarities between genes. Since the identified clusters will be used for classification, the similarity should consider the following: 1. The genes within the same cluster should be related to each other; 2. The genes should be discriminative by themselves; 3. The discriminative power should be increased by combining the LLRs of the genes within the same cluster. So we define the similarity

between genes g_i and g_k as follows:

$$s(i, k) = \begin{cases} t_k + \min\{t_{ik} - t_i, t_{ik} - t_k\} - \alpha \cdot |t_i - t_k|, & d(i, k) \leq 2 \\ -\text{Inf}, & d(i, k) > 2 \end{cases} \quad (4.4)$$

where t_k and t_i are the t -scores of the LLRs [34] of the two genes, t_{ik} is the t -score of the combined LLRs of g_i and g_k , i.e. the activity score of the subnetwork consists of only g_i and g_k . $d(i, k)$ is the length of the shortest path between g_i and g_k . So we only consider the first-order and second-order neighbors for each gene in the PPI network. When $d(i, k) < 2$, the similarity $s(i, k)$ consists of three terms. The first term of the similarity $s(i, k)$ is simply the discriminative power of g_k . The second term, $\min\{t_{ik} - t_i, t_{ik} - t_k\}$, measures the improvement in discriminative power by combining g_i and g_k . Since we only prefer the situation that the discriminative power of the combined LLRs t_{ik} is larger than both t_i and t_k , therefore the smaller one between $t_{ik} - t_i$ and $t_{ik} - t_k$ is chosen. The third term indicates penalty from the difference between the discriminative power of g_i and g_k . The parameter α , $0 \leq \alpha \leq 1$, is used to control the magnitude of this penalty term.

By using the above definition of the similarity between genes, the input similarity matrix is asymmetric. The only difference between $s(i, k)$ and $s(k, i)$ is on the first term, which means that for a pair of genes in the PPI network, the one has high discriminative power is more likely to be the exemplar.

The second input is the preference $s(k, k)$ for each gene g_k . In this application, we set the preferences to a common value c such that only 1% of the similarities between genes are larger than c . The identical value for all the preferences means that all the genes are initially equally suitable as exemplars and the set of exemplar emerges purely through competition. However, even if the preference for each gene is set to be a common value c , we can still affect the number clusters by changing c .

As shown in [38], larger c will lead to more clusters since all the data points tend to become exemplars.

B. Results and Discussion

1. The Identified Gene Clusters Improve Classification Performance Significantly

We identified the gene clusters for each dataset based on three different values of α , $\alpha = 0.2, 0.5$, and 0.8 . For each α , we first compute the similarities between genes (See Methods). The preference $s(k, k)$ for each gene in the network to a common value which is only smaller than 1% of the similarities between genes. Only the 50 clusters with highest discriminative powers are chosen for each dataset. The size of the identified gene clusters are very sensitive to the value of α . Recall the definition of the similarity between two genes, the third term $\alpha \cdot |t_i - t_k|$ is used to penalize the different of discriminative power between two genes. When α increases, only the genes with very close t -score are likely to fall into the same cluster, therefore the size of the identified cluster will decrease. Table III shows the average size of the 50 gene clusters for each dataset identified using different α . For comparison, the average size of subnetworks identified using the method proposed in Chapter III is also shown in Table III

Then, we perform the same within-dataset experiments and cross-dataset experiments as in Chapter III. The final performance is reported as the average AUC over 100 iterations of 10-fold cross-validation experiments. Figure 15 shows the performance of gene clusters identified using different α as well as the subnetworks identified using the method proposed in Chapter III with $\theta = 8$. When $\alpha = 0.5$, the identified clusters have moderate size with an average around 13 for the USA dataset and 14 for the Netherlands dataset. The performance of the identified clusters are signifi-

Table III. Average size of the identified gene clusters.

	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$	Dynamic Programming
USA	30.64	12.72	8.48	20.7
Netherlands	34	14.42	9.48	19.52

cantly better than the subnetworks identified based on dynamic programming in both within-dataset and cross-dataset experiment. When $\alpha = 0.2$, the magnitude of the penalty term will be smaller, which allows genes with different t -scores to gather. So, the average size of the gene clusters are much bigger than $\alpha = 0.5$ and 0.8 . The identified gene clusters achieve extremely good performance in within-dataset experiments and cross-dataset experiments labeled "Netherlands-USA", and only the performance of the cross-dataset experiment labeled "USA-Netherlands" is a little bit lower than the performance of the subnetworks. When $\alpha = 0.8$, the penalty term in the definition gets very big, therefore the identified clusters only contain genes with very similar t -scores, which leads to reduced cluster sizes compared to $\alpha = 0.2$ and 0.5 . The performance of the identified gene clusters is only slightly better than subnetworks in the within-dataset experiment based on the USA dataset and the cross-dataset experiment labeled "USA-Netherlands". For the other two experiments, subnetworks identified using the method proposed in Chapter III achieve better performance. A possible reason is, genes with significant discriminative power, which should be include in a certain cluster, may be excluded because of the heavy penalty from the difference of t -scores.

In order to show the impact of α on the identified clusters, we compared the 50 clusters identified based on each dataset using $\alpha = 0.2, 0.5$ and 0.8 . Figure 16 shows the statistics of the identified gene clusters, including the total numbers of

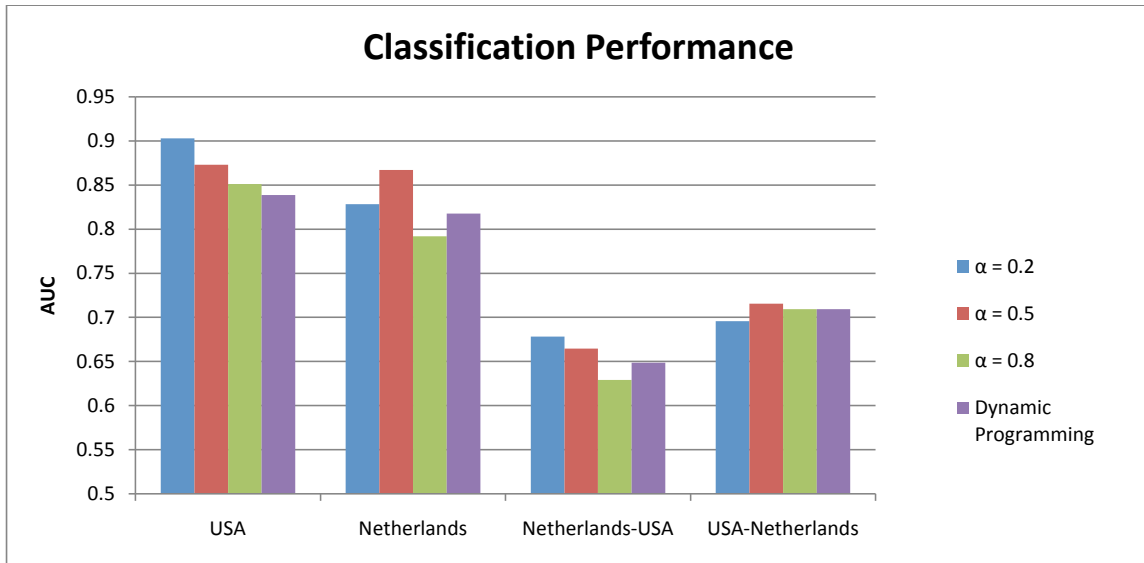


Fig. 15. Classification performance of gene clusters identified using different α .

The bar charts show the average AUC of different classifiers based on the gene clusters identified using $\alpha = 0.2, 0.5, 0.8$ and subnetwork markers found by the method proposed in Chapter III. Results of the within-dataset experiments based on the USA and Netherlands dataset are shown in the two bar charts on the left. The two bar charts on the right show the results of the cross-dataset experiments, where markers were identified based on the first dataset and tested based on the second dataset.

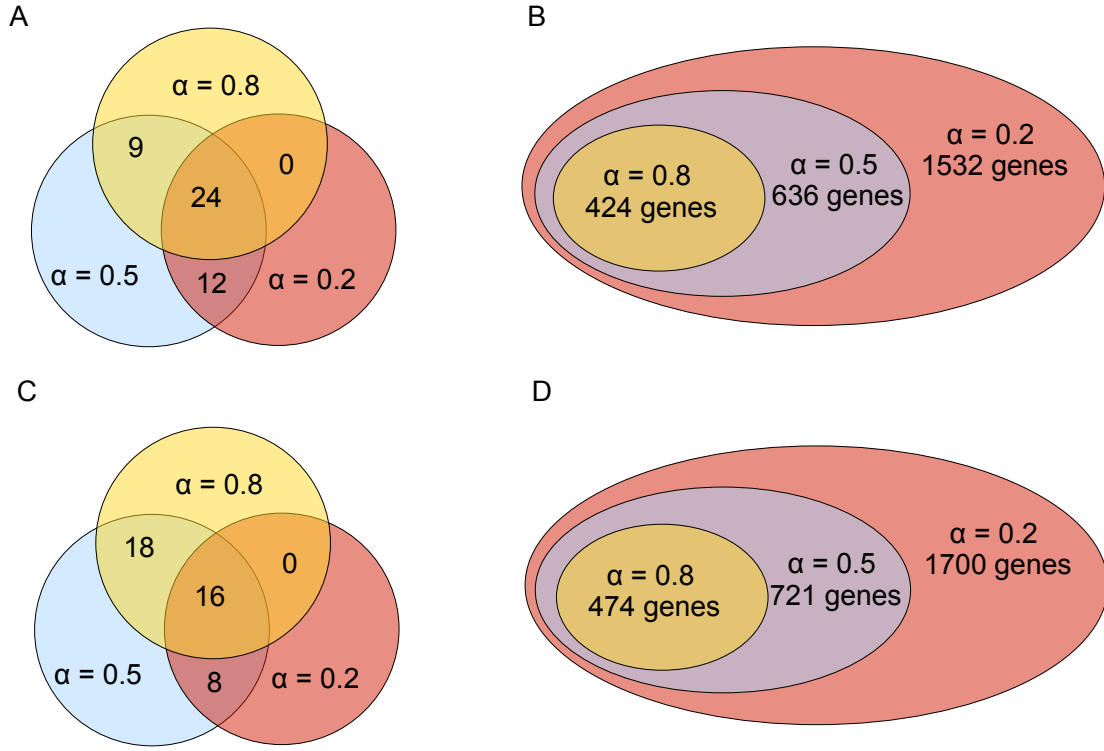


Fig. 16. Statistics of the gene clusters identified using different α .

Figure A, B show the statistics of the gene clusters identified based on the USA dataset. Figure C, D show the statistics of the gene clusters identified based on the Netherlands dataset.

genes included, the overlap between the identified gene clusters using different α , the overlap between identified exemplars. For both datasets, the number of genes included in the identified clusters decreases as α increases. Over 95% of the genes included in the identified clusters using a smaller α are also found in the clusters identified using larger α . Besides, there are significant overlaps (around 60%) between the sets of exemplars for different α , which usually contain powerful representative genes. When α gets bigger, similarities between genes will decrease generally. Therefore, more and more genes are removed from the clusters centered at those representative genes because of the decreased similarities.

2. Affinity Propagation Finds Better Combination of Genes

We further compared the identified gene clusters with the subnetworks that identified using the method proposed in Chapter III in terms of their discriminative power and the genes included.

Figure 17 shows the discriminative power of the 50 gene clusters identified using $\alpha = 0.5$ and the subnetworks identified using $\theta = 8$. It seems that the discriminative power of the gene clusters is lower than that of the subnetworks. Interestingly, the classification performance shown in Figure 15 supports that the identified gene clusters are better diagnostic markers compared to subnetworks. Although subnetworks identified using the method proposed in Chapter III have higher discriminative power than gene clusters found using affinity propagation individually, it doesn't necessarily mean that the classification performance based on the cooperation of a group of such subnetworks is better than the performance of a group of such gene clusters.

Table IV shows the statistics of genes included in the 50 subnetworks and the 50 gene clusters. For both datasets, there are significant overlaps between the diagnostic markers that found using two different methods. Around 70% of the genes included in the gene clusters identified using affinity propagation are found in the subnetworks using the previous method. Besides, we noticed that over 85% of the genes in the top 10 subnetworks for both datasets can be found in the 50 gene clusters.

C. Conclusions

In this chapter, we proposed to use a clustering algorithm, named affinity propagation, to cluster genes based on the protein-protein interaction(PPI) network. This algorithm takes the input as the measures of similarity between pairs of data points, and makes decision by exchanging messages between data points. Based on our def-

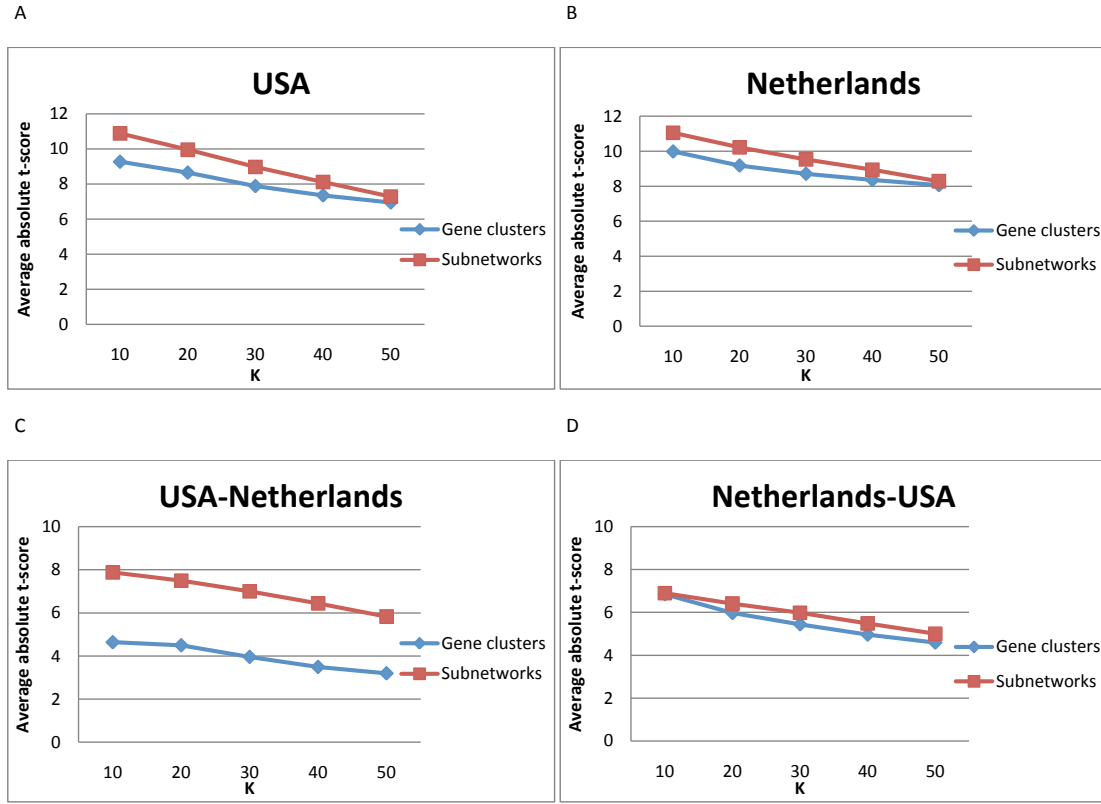


Fig. 17. Discriminative power of the identified gene clusters and subnetworks.

Figure A, B show the within-dataset experiment results. Figure C, D show the cross-dataset experiment results.

Table IV. Statistics of the gene clusters identified using affinity propagation and the subnetworks identified using the method proposed in Chapter III.

Dataset		Number of genes	Size of overlaps
USA	Gene clusters	636	438
	Subnetworks	1035	
Netherlands	Gene clusters	721	474
	Subnetworks	976	

The 50 gene clusters are identified using $\alpha = 0.5$ for both datasets. The 50 subnetworks are identified with $\theta = 8$.

initiation of similarity between pairs of genes that are connected in the network, this algorithm finds the set of exemplars as well as their corresponding clusters, which can be used for disease classification. One main advantage of the proposed method compared to the previous methods is that there is no pre-selection procedure included in this algorithm. Affinity propagation considers all the genes in the network as exemplars initially. The set of representative genes emerges gradually in the process of affinity propagation. Moreover, affinity propagation finds non-overlapping gene clusters simultaneously rather than removing the identified subnetworks from the whole network. Therefore, affinity propagation is more likely to find the appropriate combination of genes, which can serve as better diagnostic markers for classification. The cross-validation results based on two independent breast cancer datasets show that the gene clusters identified using affinity propagation can yield more accurate and reliable classification performance compared to the subnetworks identified in Chapter III.

CHAPTER V

CONCLUSION

In this thesis, we first proposed a probabilistic model for pathway and subnetwork activity in chapter II. The simulation results based on two independent breast cancer datasets show that the proposed inference method is more efficient compared to the previous inference methods. In chapter III and chapter IV, we introduced two methods for identifying subnetworks or gene clusters based on a large protein-protein interaction network. The identified subnetworks and gene clusters were tested based on the same breast cancer datasets used in chapter II. The results show that subnetworks and gene clusters can server as diagnostic markers that yield better classification performance compared to single gene makers, pathway markers and subnetwork markers identified using a previous method.

REFERENCES

- [1] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, “Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling,” *Nature*, vol. 403, pp. 503–511, Feb 2000.
- [2] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, “Molecular classification of cancer: class discovery and class prediction by gene expression monitoring,” *Science*, vol. 286, pp. 531–537, Oct 1999.
- [3] A. Perez-Diez, A. Morgun, and N. Shulzhenko, “Microarrays for cancer diagnosis and classification,” *Adv. Exp. Med. Biol.*, vol. 593, pp. 74–85, 2007.
- [4] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub, “A molecular signature of metastasis in primary solid tumors,” *Nat. Genet.*, vol. 33, pp. 49–54, Jan 2003.
- [5] B. Efron and R. Tibshirani, “Empirical bayes methods and false discovery rates for microarrays,” *Genet. Epidemiol.*, vol. 23, pp. 70–86, Jun 2002.
- [6] P. Baldi and A. D. Long, “A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes,” *Bioinformatics*, vol. 17, pp. 509–519, Jun 2001.

- [7] T. B. Kepler, L. Crosby, and K. T. Morgan, “Normalization and analysis of DNA microarray data by self-consistency and local regression,” *Genome Biol.*, vol. 3, pp. RESEARCH0037, Jun 2002.
- [8] T. Ideker, V. Thorsson, A. F. Siegel, and L. E. Hood, “Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data,” *J. Comput. Biol.*, vol. 7, pp. 805–817, 2000.
- [9] Y. Chen, E. R. Dougherty, and M. L. Bittner, “Ratio-based decisions and the quantitative analysis of cdna microarray images,” *Journal of Biomedical Optics*, vol. 2, pp. 364–374, 1997.
- [10] L. J. van ’t Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, “Gene expression profiling predicts clinical outcome of breast cancer,” *Nature*, vol. 415, pp. 530–536, Jan 2002.
- [11] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Tantalov, M. Timmermans, M. E. Meijer-van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens, “Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer,” *Lancet*, vol. 365, pp. 671–679, 2005.
- [12] M. West, C. Blanchette, H. Dressman, E. Huang, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, J. R. Marks, and J. R. Nevins, “Predicting the clinical status of human breast cancer by using gene expression profiles,” *Proc. Natl. Acad. Sci. U.S.A.*, Sep 2001, vol. 98, pp. 11462–11467.

- [13] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?," *Bioinformatics*, vol. 20, pp. 374–380, Feb 2004.
- [14] U. M. Braga-Neto, "Fads and fallacies in the name of small-sample microarray classification.," *IEEE Signal Processing Magazine*, vol. 20, pp. 91, 2007.
- [15] E. R. Dougherty, "Small sample issues for microarray-based classification," *Comp. Funct. Genomics*, vol. 2, pp. 28–34, 2001.
- [16] A. Dupuy and R. M. Simon, "Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting," *J. Natl. Cancer Inst.*, vol. 99, pp. 147–157, Jan 2007.
- [17] L. Ein-Dor, O. Zuk, and E. Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proc. Natl. Acad. Sci. U.S.A.*, Apr 2006, vol. 103, pp. 5923–5928.
- [18] S. Michiels, S. Koscielny, and C. Hill, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *Lancet*, vol. 365, pp. 488–492, 2005.
- [19] E. E. Ntzani and J. P. Ioannidis, "Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment," *Lancet*, vol. 362, pp. 1439–1444, Nov 2003.
- [20] Dougherty ER Hua J, Tembe WD, "Performance of feature-selection methods in the classification of high-dimension data," *Pattern Recognition*, vol. 42, pp. 409–424, 2008.

- [21] A. Subramanian, P. Tamayo, V. K. feature, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles,” *Proc. Natl. Acad. Sci. U.S.A.*, Oct 2005, vol. 102, pp. 15545–15550.
- [22] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park, “Discovering statistically significant pathways in expression profiling studies,” *Proc. Natl. Acad. Sci. U.S.A.*, Sep 2005, vol. 102, pp. 13544–13549.
- [23] A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M. B. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck, J. A. Olson, J. R. Marks, H. K. Dressman, M. West, and J. R. Nevins, “Oncogenic pathway signatures in human cancers as a guide to targeted therapies,” *Nature*, vol. 439, pp. 353–357, Jan 2006.
- [24] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium,” *Nat. Genet.*, vol. 25, pp. 25–29, May 2000.
- [25] Z. Guo, T. Zhang, X. Li, Q. Wang, J. Xu, H. Yu, J. Zhu, H. Wang, C. Wang, E. J. Topol, Q. Wang, and S. Rao, “Towards precise classification of cancers based on robust gene functional expression profiles,” *BMC Bioinformatics*, vol. 6, pp. 58, 2005.
- [26] E. Lee, H. Y. Chuang, J. W. Kim, T. Ideker, and D. Lee, “Inferring pathway

- activity toward precise disease classification,” *PLoS Comput. Biol.*, vol. 4, pp. e1000217, Nov 2008.
- [27] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J. P. Vert, “Classification of microarray data using gene networks,” *BMC Bioinformatics*, vol. 8, pp. 35, 2007.
- [28] J. Tomfohr, J. Lu, and T. B. Kepler, “Pathway level analysis of gene expression using singular value decomposition,” *BMC Bioinformatics*, vol. 6, pp. 225, 2005.
- [29] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker, “Network-based classification of breast cancer metastasis,” *Mol. Syst. Biol.*, vol. 3, pp. 140, 2007.
- [30] M. Kanehisa, S. Goto, S. Kawashima, and A. Nakaya, “The KEGG databases at GenomeNet,” *Nucleic Acids Res.*, vol. 30, pp. 42–46, Jan 2002.
- [31] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor, and B. R. Conklin, “GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways,” *Nat. Genet.*, vol. 31, pp. 19–20, May 2002.
- [32] T. Fawcett, “An introduction to ROC analysis,” *Patt Recog Letters*, vol. 27, pp. 861–874, Jun 2006.
- [33] S. Efroni, C. F. Schaefer, and K. H. Buetow, “Identification of key processes underlying cancer phenotypes using biologic pathway analysis,” *PLoS ONE*, vol. 2, pp. e425, 2007.
- [34] J. Su, B. J. Yoon, and E. R. Dougherty, “Accurate and reliable cancer classification based on probabilistic inference of pathway activity,” *PLoS ONE*, vol. 4, pp. e8161, 2009.

- [35] G. F. Berriz, J. E. Beaver, C. Cenik, M. Tasan, and F. P. Roth, “Next generation software for functional trend analysis,” *Bioinformatics*, vol. 25, pp. 3043–3044, Nov 2009.
- [36] H. C. Mak, M. Daly, B. Gruebel, and T. Ideker, “CellCircuits: a database of protein network models,” *Nucleic Acids Res.*, vol. 35, pp. D538–545, Jan 2007.
- [37] J. Su and B.-J. Yoon, “Identifying reliable subnetwork markers in protein-protein interaction network for classification of breast cancer metastasis,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010.
- [38] B. J. Frey and D. Dueck, “Clustering by passing messages between data points,” *Science*, vol. 315, pp. 972–976, Feb 2007.

VITA

Junjie Su received his B.Eng. degree in automation from Tsinghua University, China, in 2008. He graduated with his M.S. in electrical engineering from Texas A&M University in December 2010. His research interests include systems biology and statistical pattern recognition.

Junjie Su may be reached at 214 Zachry Engineering Center, TAMU 3128, College Station, TX, 77843-3128. His email is tracysjj@gmail.com.

The typist for this thesis was Junjie Su.